Thèse Université Lumière LYON II
Ecole Doctorale de Sciences Cognitives

Traitement de la Prosodie par un Réseau Récurrent Temporel :

Jean-Marc BLANC
Sciences Cognitives – Mention Informatique
Directeur de Thèse Dr. Peter Ford DOMINEY
Institut des Sciences Cognitives
2 Février 2005

Membres du Jury Pr. Harriet JISA Dr. Axel CLEEREMANS Dr. Yves BURNOD Dr. François PELLEGRINO Dr. Peter Ford DOMINEY Rapporteurs Dr. Axel CLEEREMANS Dr. Daniel HIRST

Table des matières

A Louis Martine Nathalie Loïc

Remerciements

Si vous lisez les quelques pages qui suivent, vous aurez probablement le sentiment qu'une thèse est le travail d'une personne isolée, il n'en est rien et il revient à ces lignes de rétablir la vérité.

Mon grand-père et ma mère sont sans doute les premiers à l'origine de cette aventure et ils y contribuent encore. Mais la suite aurait été tout autre si je n'avais rencontré Peter, qui m'a guidé et a toujours été confiant dans mes idées. Je tiens à les remercier lui et sa femme tout particulièrement parce qu'ils ne m'ont jamais imposé de limites inatteignables ou intransgressibles.

C'est lui encore qui m'a permis de rencontrer la plupart des personnes qui ont contribué à ce mémoire : tout d'abord, durant mon D.E.A Gérard BAILLY m'a initié à la prosodie et a toujours fourni des commentaires utiles lors de ma thèse, Jean-Luc SCHWARTZ a été rapporteur de mon mémoire de stage, enfin François PELLEGRINO m'a introduit à l'Identification Automatique des Langues, et m'a apporté le soutien de la région Rhônes-Alpes à travers le projet émergence. Ensuite, j'ai eu l'honneur de travailler avec Morten CHRISTIANSEN sur des thématiques assez proches de celles développées dans cette thèse.

Je leur suis reconnaissant de m'avoir faire connaître leurs domaines de recherches respectifs que j'espère avoir étudié au mieux dans cet ouvrage. Je tiens également à témoigner de l'aide précieuse des doctorants, maintenant post-doctorants ou chargés de recherche, avec qui j'ai pu travailler. Jérôme FARINAS a été d'un brillant secours pour mes questions concernant le Traitement Automatique des Langues. Franck RAMUS a fourni une partie du matériel parole employée dans cette thèse. Olivier CROUZET m'a judicieusement conseillé sur les outils pour traiter la parole et m'a éclairé sur la psycholinguistique, j'exprime aussi toute ma gratitude à Christelle DODANE qui m'a transmis une partie de son savoir sur la prosodie, et avec qui j'ai pris conscience des difficultés qu'il pouvait exister pour communiquer entre deux cultures scientifiques distinctes.

Grâce à eux, je peux enfin remercierceux qui s'intéressent à mes recherches : Harriet JISA qui a des fins comparables aux miennes, mais avec des moyens différents, Axel CLEEREMANS, qui a su montrer l'intérêt des réseaux de neurones pour la psychologie, Yves BURNOD qui crée des ponts entre le monde des neurosciences et le traitement de la parole, ainsi que François PELLEGRINO et Peter Ford DOMINEY qui ont accepté de constituer mon jury de thèse.

Je remercie très chaleureusement Daniel HIRST et Axel CLEEREMANS pour avoir répondu présent à la dernière minute et consacrer une partie de leur temps pour relire ma prose.

Un grand merci pour toute l'équipe scientifique, technique et administrative de l'Institut des Sciences Cognitives (Patrice, Marc, Sabine, Anne, Carole, Hassen, Belkacem, Yves, Guy, Ira, Anne, Nadine, Emmanuelle, Sylvie, Jean-Baptiste). Je n'oublie pas non plus de remercier les pauses de l'institut, dont l'ambiance est assurée chaque jour par Hyung, Flavie, Fabrice, Nathalie, Lewis, Zoltan, Thomas, Michel, Sonia, Nelly, Wafa, David, David, Yannick, Emmanuelle, Nadia, James, Aurélie, Stéphanie, Jean-Yves, Nausicaa, Alexandre, Thierry.

Babaga Brrffff Loïc pour m'inspirer de nouveaux sujets de recherches. Enfin ma dernière pensée va à celle qui a « subit la thèse » du début jusqu' à la fin, et qui a su rester près de moi pendant tout ce temps, et être mon premier relecteur anonyme.

Résumé

Traitement de la Prosodie par un Réseau Récurrent Temporel Un Cadre Unifié pour l'Identification Automatique des Langues, des Attitudes Prosodiques, et des Catégories Lexicales.

La prosodie nous est directement accessible lorsque nous écoutons une langue étrangère. Quel mécanisme peut traiter la prosodie de la parole ? Un réseau récurrent temporel (TRN) vérifiant des études neurophysiologiques sur l'apprentissage de séquences par le primate a été testé pour l'identification de contours prosodiques définis sur différentes échelles réparties sur un continuum temporel.

Le rythme d'une langue peut être caractérisé globalement. Le réseau TRN identifie cinq langues européennes à partir d'un cochléogramme (65%). En employant la fréquence fondamentale, le réseau reconnaît six attitudes (modes syntaxiques et émotions) et distingue les mots de fonction et de contenu, deux catégories lexicales, à l'origine de l'amorçage de la syntaxe. Lorsque le modèle est altéré de façon à amoindrir sa sensibilité temporelle, cette catégorisation ne s'effectue plus et le profil des réponses à deux tâches de traitement auditif rapide est semblable à celui d'enfants ayant des troubles spécifiques du langage, en particulier pour la syntaxe. En outre, ce modèle réplique deux discriminations prosodiques réalisées par des nouveau-nés : les langues sont distinguées en fonction de leur classe rythmique et les mots de fonction se distinguent des mots de contenu.

En résumé, le réseau TRN accomplit trois tâches dans lesquelles la prosodie se définit entre un domaine global (une langue) et local (un mot) : Identification Automatique des Langues et des Attitudes Prosodiques ; Discrimination des mots de Fonction et de Contenu. Enfin, des troubles auditifs pour le traitement d'événement rapide et des troubles de la syntaxe peuvent être simulés par une déformation de la sensibilité temporelle du réseau.

Abstract

Prosody processing by a Temporal Recurrent Network A common framework for Automatic Identification of Languages, Prosodic Attitudes, and Lexical Categories.

Prosody is directly accessible to us when hearing a foreign language. What is the mechanism implicated in the processing of speech prosody ? A Temporal Recurrent Network (TRN) inspired by neurophysiologic studies for sequences learning by primates has been tested for the identification of prosodic contours.

Language rhythm can be globally defined. The TRN identify five European languages (50%) based on the automatic segmentation of speech in consonants and vowels, but also with a cochleogram (65%).

With the fundamental frequency, the network identifies six prosodic attitudes (syntactic modes and emotions) and distinguishes content from function words, two lexical categories that could bootstrap syntax. When the model is distorted in order to reduce its temporal sensitivity, this categorization could not be realized, and the pattern of response to two rapid auditory processing tasks resembles that of children with Specific Language Impairment, in particular for syntax.

In addition, this model replicates two experiments of prosodic discrimination realized by new-borns: languages are distinguished according to their class rhythm and function words are discriminated from content words.

In brief, the TRN accomplishes three tasks where prosody is defined on different temporal domains: from language (global field) to word (local field): Automatic Identification of Languages, and prosodic attitudes; Discrimination of content and function words. Finally auditory and language specific troubles could be simulated when the temporal sensitivity of the network is weaken.

Introduction

« Tous les sens sont connectés d'une manière ou d'une autre au cerveau ; en conséquence, ils sont incapables d'agir si le cerveau est dérangé ou qu'il change de position car cet organe bloque les passages par lesquels les sens agissent. » Théophraste reprenant les propos enseignés par Alcméon, De Sensibus .

I.Les mécanismes cognitifs

I.1.L'étude du cerveau

Depuis l'antiquité Grecque, l'homme s'interroge sur les mécanismes qui sous-tendent sa perception du monde extérieur. Ses sensations se révèlent comme fonction des organes des sens *et* du cerveau. Hippocrate avancera (dans *la maladie sacrée*) que le cerveau est responsable de la totalité des processus cognitifs, affectifs et conatifs (Marshall, 2002). Comment alors étudier le fonctionnement du cerveau, et plus particulièrement sa capacité à traiter les informations transmises par nos sens ?

Ce problème est rendu d'autant plus ardu par la multiplicité des tâches que le cerveau est susceptible d'effectuer. Il est encore impossible, à l'heure actuelle, de pouvoir étudier le système dans sa totalité. Une solution consiste donc à réduire le problème à

des tâches extrêmement codifiées, pour pouvoir obtenir un modèle satisfaisant.

Quelques années auparavant, l'ingénierie moderne a cru pouvoir maîtriser rapidement les problèmes auxquels l'espèce humaine est confrontée tous les jours sans même en avoir conscience. Ce fût le premier écueil dans la courte vie de l'Intelligence Artificielle. Toutes ces questions sont relatives au traitement de l'information transmise par nos cinq sens à notre cerveau, et à la production d'une réponse adaptée à la perception de notre environnement.

L'apprentissage de la marche, l'identification des odeurs, la maîtrise des saveurs, la reconnaissance des objets et des scènes et l'acquisition du langage sont des problèmes d'une étrange complexité, nécessitant l'appréhension de dimensions différentes dans l'espace et le temps, alors que celles-ci sont traitées par un unique organisme. Sous cet angle, il est alors particulièrement compliqué de réduire notre esprit à un modèle ou une machine théorique capable de répliquer les mêmes processus, qui nous sont si courants.

Malgré ces obstacles, cette quête du Graal reste l'objectif majeur des neurosciences.

I.2.Les neurosciences

L'un des premiers buts des neurosciences consiste à relier un réseau de zones du cerveau avec une ou plusieurs opérations. Ainsi, Broca a été le premier à associer de façon explicite un trouble du langage articulé à la lésion d'une zone du cerveau, dénommée dès lors « aire de Broca », donnant ainsi le jour à la neuropsychologie. De même, les aires auditives primaires sont systématiquement situées dans les circonvolutions de Heschl, enfouies dans la scissure de Sylvius.

Ces deux exemples donnent une illustration de phénomènes généraux, comme la production du langage, ou le traitement auditif. Mais, des fonctions plus spécifiques peuvent également être localisées, comme l'extraction du contour de la hauteur d'un son [1] (Peretz, 2000). Par conséquent, les neurosciences établissent une carte fonctionnelle du cerveau. Elles tentent maintenant d'expliquer le fonctionnement de ces mécanismes, en fournissant des indications sur l'architecture et les propriétés de réseaux de neurones spécifiques.

Un enjeu complémentaire est la recherche des communications effectuées entre ces différents mécanismes. Ainsi, pour comprendre le fonctionnement d'une tâche de haut niveau, comme la parole, il faut définir la chaîne des différents mécanismes, comme la perception de la hauteur, et le traitement de la parole, de façon à percevoir le contenu informatif d'une phrase et l'émotion qui est véhiculée.

Pour répondre à ces objectifs, les neurosciences observent le cerveau à l'aide de dispositifs comme l'Imagerie par Résonance Magnétique fonctionnelle (IRMf) ou la Tomographie par Emission de Positrons (TEP). Ces méthodes façonnent une cartographie du flux sanguin cérébral, pendant des tâches effectuées par un sujet. L'étude de ces cartes permet alors de conclure sur les régions recrutées pour l'exécution d'un processus. Ces techniques nécessitent une fenêtre d'intégration temporelle de

[1] Hémisphère droit ; région du gyrus temporal supérieur et région frontale.

quelques secondes, ce qui rend difficile l'examen des processus qui se déroulent en moins de quelques centièmes de millisecondes. Une tâche cognitive se décompose en opérations mentales élémentaires, d'une durée de dix à cent millisecondes. L'examen des champs électromagnétiques (électroencéphalogramme ou EEG) avec des techniques telles que les potentiels évoqués (ERPs) décrivent les événements à la milliseconde près. Ces techniques abordent un autre sujet brûlant de la recherche en sciences cognitives : l'intégration du temps dans les processus cognitifs.

II.Le temps

L'information sensorielle doit être interprétée pour permettre une perception de l'environnement. Quel que ce soit le sens mis en jeu, les événements se succèdent dans le temps. Par conséquent, l'élaboration d'une représentation mentale s'avère indispensable pour percevoir leur structure. Traiter l'information sensorielle est un art du temps qui nécessite donc pour être comprise, ***« non seulement un acte de mémorisation, mais encore un effort constant pour relier le passé au présent, relation qui est en définitive d'ordre intellectuel et non perceptif. »*** (Imberty, 1969, p.115) [2] .

II.1.L'intégration du temps au sein des structures cérébrales

La perception de notre environnement est basée sur des motifs spatio-temporels d'activité neuronale, qui sont le résultat des entrées sensorielles. En décodant ces motifs, le cerveau interprète ce qui est perçu.

Il est important de séparer les dimensions temporelles et spatiales, l'une de l'autre. Dans le cas de la vision, les lignes horizontales et verticales activent différentes régions spatiales de la rétine. En ce qui concerne l'audition, des tons de fréquences distinctes activent différentes régions spatiales de la cochlée. Ainsi, au cours du temps, le même groupe de cellules répond à deux stimuli ayant même fréquence (ou même orientation spatiale). En conséquence, il faut donc un mécanisme spécifique (i.e. un autre groupe de cellules) pour répondre sélectivement à la durée des événements. Ceci peut s'effectuer par le biais d'une transformation de la dimension temporelle vers la dimension spatiale, qui pourrait être opérée par la machine cérébrale.

Cette première opération permet donc d'obtenir les événements, extraits du flux temporel des signaux sensoriels. L'ordre sériel de ces événements et des actions est critique pour l'étude de la cognition et du comportement. En abordant cet aspect, il y a plus de cinquante ans, Lashley (1951) postulait que le cerveau analysait et contrôlait l'ordre sériel, en utilisant les motifs spatiaux d'activités neuronales (les idées). Pour contrôler les séquences d'actions, ces motifs spatiaux devaient être transformés en

[2] Cette phrase s'adressait principalement à la perception auditive, qui est sans doute particulièrement marqué par le temps, mais toute perception s'effectue dans le temps.

actions dans le temps, par un procédé qu'il a dénommé « syntaxe » dans la formation du langage par les idées (Beiser et Houk, 1998).

Des enregistrements de cellules ont été effectués chez le primate en train d'exécuter des tâches sensori-motrice de reproduction de séquences (Barone et Joseph, 1989). Les réponses qui sont induites par les séquences pendant la période précédant la réponse motrice peuvent représenter la conversion des séquences temporelles des entrées sensorielles en un motif spatial d'activité neuronale.

Un tel motif peut représenter les commandes de la conversion inverse, c'est-à-dire d'un motif spatial vers un motif temporel, qui sera à l'origine du mouvement. Cette étude prouve l'existence d'un mécanisme effectuant ces conversions, donnant ainsi crédit aux idées qu'avait exprimées Lashley en 1951 (Barone et Joseph, 1989).

Notre système nerveux traite les informations sensorielles au cours du temps, ainsi que leurs durées. Il faut maintenant pouvoir tester si une carte fonctionnelle dédiée à ce traitement peut être utilisée dans une implémentation informatique pour résoudre les problèmes liés à la dimension temporelle, traités par ces fonctions du cerveau.

II.2.Les neurosciences computationnelles

L'ambition des neurosciences computationnelles est de pouvoir simuler les réseaux de neurones biologiques par le biais des machines et comprendre ainsi les mécanismes neuronaux, qui permettent de transformer l'activité des neurones en fonction corticale de haut niveau. Les modèles employés s'inspirent des détails anatomiques et physiologiques, ainsi que des données comportementales pour reproduire des expériences biologiques.

Les réseaux de neurones artificiels ont recours à la modélisation d'unités simples (les neurones artificiels) qui dupliquent les propriétés de leurs cousins biologiques. Ce champ de recherche se doit donc d'une part, d'extraire le maximum de paramètres pour caractériser le plus fidèlement les neurones et les canaux qui les relient (connexions et axones), et d'autre part, de comprendre les liens structurels et fonctionnels entre les divers réseaux de neurones.

La majorité des travaux en Intelligence Artificielle s'appuie sur des modèles symboliques de réseaux de neurones, obéissant à diverses règles d'apprentissage permettant l'ajustement du poids des connexions. De plus en plus, les neurosciences computationnelles utilisent des modèles en temps continu, qui représentent soit les variations du nombre moyen de décharge, soit l'évolution temporelle du potentiel de membrane (Grethe et Arbib, 2001).

Des structures cérébrales ont inspiré des modèles informatiques (Grethe et Arbib, 2001) :

- **Interactions pariétal - aire prémotrice, pour le contrôle de la saisie (grasping) :** Fondées sur des données empiriques et des observations d'expériences PET chez l'être humain.

- **Les glandes de bases :** Leurs rôles dans le contrôle des saccades et des bras, ainsi que pour le traitement séquentiel, et ses effets sur le comportement des sujets parkinsoniens.
- **Cervelet :** Conditionnement et coordination des capacités motrices.
- **Hippocampe :** Les recherches neurochimiques et neurophysiologiques sur la potentialisation à long terme (LTP) ont permis d'obtenir une simulation de la synapse (Liaw et Berger, 1996 ; 1998). Des relations entre l'hippocampe et de nombreuses autres régions du cerveau ont été mises à jour.
- **Cortex préfrontal, connexions récurrentes cortico-corticales, synapses cortico-striatales :** Ces structures constituent un modèle de Réseau Récurrent Temporel (TRN) construit pour apprendre des séquences sensori-motrices (Dominey, Arbib et Joseph, 1995). Ce réseau repose sur des données comportementales et des enregistrements neurophysiologiques chez le primate non-humain, qui suggèrent que le cortex préfrontal encode à la fois la position spatiale et l'ordre séquentiel des évènements. Il permet de traiter la structure temporelle, en se libérant de certaines contraintes liées à l'apprentissage chez ce type de réseau.

Nous venons de donner un aperçu des moyens de modèlisation des systèmes cérébrales dédiés au traitement temporel. Quel matériel ces réseaux peuvent-ils manipuler ?

III.La parole

III.1.Sa spécificité

A la suite d'un célèbre débat entre Chomsky et Skinner, Al Liberman et ses collaborateurs du laboratoire de Haskins ont commencé à explorer en détail les mécanismes qui sous-tendent la perception de la parole chez l'être humain (Hauser, 2002). L' « organe du langage » et sa capacité à produire des structures syntaxiques semble propre aux êtres humains. Celui-ci aurait évolué pour des raisons indépendantes de sa nécessité dans la communication.

Dès lors des scientifiques ont exhibé des aptitudes à manipuler les éléments du langage chez d'autres races animales. Par exemple, les grands singes peuvent enchaîner des symboles pour former ou comprendre des phrases. Des petits singes sauvages utilisent des vocalisations pour désigner des objets ou des événements de leur environnement.

Plus récemment, les efforts pour mettre en lumière ces mécanismes de traitement propre au langage se sont concentrés sur les processus traitant les signaux de parole. Le premier traitement abordé démontre l'existence d'une perception catégorielle des stimuli verbaux chez le nourrisson humain et chez l'animal non humain (les primates, mais aussi les oiseaux). Ainsi, un mécanisme de perception catégorielle aurait évolué avant

l'apparition du langage.

Cependant, Kuhl (1991) montre que l'adulte et le nourrisson humain, mais pas le singe Rhésus, perçoivent une distinction entre les bons et les mauvais exemplaires d'une même classe phonétique. Les bons exemples fonctionnent comme des aimants perceptifs fixant la catégorie (prototypes), et perturbent la distinction des sons proches de ce prototype. En outre, l'expérience de l'environnement linguistique influe sur leurs formations. Par conséquent chaque communauté linguistique possède des prototypes ajustés à la phonologie particulière de sa langue naturelle. Toutefois, le sansonnet montre aussi cet effet d'aimant perceptif (Kluender et coll., 1998).

Ces études semblent indiquer que l'être humain a hérité des animaux un ensemble de mécanismes perceptifs généraux pour écouter la parole. Cependant, l'être humain reste unique lors de l'utilisation de la récursivité, qui lui permet de combiner des éléments en un nombre infini d'expressions affectées de sens, donnant ainsi naissance au langage. Comment l'être humain parvient-il à maîtriser le langage ?

III.2.Son acquisition

L'apprentissage d'une langue implique de maîtriser certaines dimensions propres à la parole :

- **La phonétique :** le répertoire des sons d'une langue ;
- **La phonologie :** les structures syllabiques, accentuelles, la réalisation effective des sons d'une langue ;
- **La morphologie :** les règles qui définissent les inflexions ;
- **Le lexique :** le répertoire des mots et leur définition (la sémantique).
- Dans le domaine de la perception du langage parlé chez le nourrisson, les premières questions portaient sur le répertoire des aptitudes du nourrisson. Les travaux de P. Eimas et de ses collaborateurs avaient démontré que le nourrisson pouvait percevoir des contrastes phonétiques bien avant de n'en avoir jamais produits (Jusczyk, 2002). Soudain, les chercheurs intéressés par l'acquisition du langage commencèrent à prendre au sérieux la possibilité que le nourrisson y soit biologiquement préparé et qu'il possède des capacités d'acquisition spécialisées.

Quelle est l'étendue de ces capacités à la naissance ? J. Mehler a démontré que le nourrisson discrimine mieux un contraste apparaissant dans des syllabes bien formées, que dans des syllabes qui n'appartiennent pas à sa langue maternelle (Mehler et Bertoncini, 1981). Kuhl, Williams, Lacerda, Stevens, et Lindblom (1992) ont émis l'idée que le nourrisson commence à développer des catégories pour les voyelles de sa langue maternelle dès l'âge de six mois. Ensuite, le nourrisson acquiert des informations sur l'enchaînement des phonèmes (contraintes phonotactiques) qui peuvent apparaître dans sa langue maternelle. Les nourrissons de 9 mois sont sensibles à la distribution typique de l'accent tonique en Anglais (Jusczyk, Cutler et Redanz, 1993). Jusczyk et Aslin (1995) ont également montré les aptitudes de segmentation du signal de parole, pour des bébés

de 7 mois et demi. Effectivement ils sont capables de reconnaître des mots cibles présentés isolément, alors que ceux-ci leur avaient été présentés à l'intérieur d'une phrase.

Aujourd'hui, les chercheurs s'intéressent particulièrement à l'acquisition de la grammaire d'une langue, notamment à travers les informations contenues dans le signal de parole (Morgan et Demuth, 1996).

III.3.Sa syntaxe

Dans son approche théorique de l'acquisition du langage, Chomsky (1965) note que l'enfant doit avoir deux techniques distinctes pour représenter le signal de parole et la structure des informations contenues dans ce signal. Cette seconde technique consiste en un ensemble de règles permettant d'établir une structure hiérarchique à partir d'une phrase du langage. Les deux termes syntaxe et grammaire sont employés pour décrire cette fonction. Cette organisation reflète alors la structure qui est inhérente à l'information communiquée, et donne un éclairage particulier à la scène décrite.

Les théories de la syntaxe décrivent la structure des phrases en terme de constituants (nom, verbe, adjectif, préposition, …), de fonctions grammaticales (sujet, objet, complément …), de cas (nominatif, accusatif, datif, etc.) pour pouvoir appréhender la syntaxe, il faut donc apprendre les catégories syntaxiques des mots, et l'organisation de ces catégories entre elles. Se basant sur la segmentation de la parole en mots, le sens de ces mots, ainsi que le sens global des phrases, l'enfant peut alors essayer de déduire les règles grammaticales qui gouvernent les phrases qu'il entend.

Les nourrissons se trouvent confrontés en même temps à ces deux problèmes d'une grande complexité : la signification des mots dans le signal de parole (le **lexique**), et la découverte de l'organisation hiérarchique entre ces mots (la **syntaxe**). La situation qui s'en suit est paradoxale, puisque chacun de ces problèmes ne semble pouvoir être résolu l'un sans l'autre. Les théories actuelles tombent souvent dans ce cercle vicieux. Ainsi le sens des verbes nécessite la connaissance de l'ordre des mots, connaissance qui requiert la signification des mots, en particulier les verbes.

Cette circularité peut être brisée si le signal de parole lui-même comprend des indices qui indiquent par exemple les différentes catégories lexicales ou l'ordre des mots. Dans ce cas, un sous-ensemble de chaque problème peut se définir indépendamment de l'autre. Cette correspondance entre des propriétés acoustiques du signal de parole et des caractéristiques linguistiques peut alors être à l'origine de différents problèmes posés par l'acquisition du langage. La perception de la parole et sa représentation contribue à l'acquisition de la syntaxe.

III.4.La prosodie : Rendez-vous du temps et de la parole

Nous venons de décrire deux des nombreuses thématiques abordées dans la recherche en sciences cognitives : le traitement des informations inscrites dans le temps par le système nerveux, et la question de l'acquisition et du traitement de la parole par les êtres

humains. Ces champs de recherches se rejoignent dans la prosodie. Celle-ci traduit entre autre la dimension temporelle de la parole, à travers son rythme, sa mélodie, son volume sonore, et son timbre.

S'il est clair que le traitement de la parole s'effectue au cours du temps, il n'est pas aussi évident que les particularités rythmiques de la parole puissent concourir à la compréhension et l'acquisition de la parole. Néanmoins, un champ de la recherche de la linguistique a pour objectif d'étudier les caractéristiques suprasegmentales, comme le rythme. Il s'agit bien sûr de la prosodie, qui définit l'intonation, les accents, les pauses, en fonction du décours temporel des phrases.

La prosodie peut s'organiser autour de trois échelles temporelles :

1. Une échelle globale, qui caractérise les la totalité des informations prosodiques pour une langue. Par exemple, le rythme permet de distinguer certaines langues.
2. Une échelle intermédiaire, qui définit des séquences de valeurs sur la durée d'une phrase. Ainsi, l'intonation distingue les différentes émotions qui peuvent être transmises par une même phrase.
3. Une échelle locale, qui comporte des caractéristiques comme les accents liés à une unité atomique de la parole, comme la syllabe.

Notre objectif est d'adapter un modèle issu des neurosciences au traitement de la prosodie, en vérifiant que ce modèle peut résoudre des tâches d'identification qui s'insèrent chacune dans une des échelles précitées. En outre, nous proposons une simulation pour étudier les problèmes soulevés si le modèle est modifié de façon à perturber le traitement de l'échelle locale.

IV.Plan de thèse

La parole est organisée par sa structure temporelle, suivant différentes échelles de temps. Concrètement, nous étudierons trois niveaux distincts, qui permettent à un auditeur de déterminer (a) la langue parlée, (b) les attitudes communicatives portée par l'intonation, (c) les catégories lexicales des mots, et finalement (d) la présence ou l'ordre de tons purs brefs. A priori, ces questions peuvent sembler disparates et sans point commun : en particulier, chacun de ces niveaux d'analyse peut nécessiter un mécanisme de traitement distinct.

Cette thèse postulera une position théorique contraire, que nous dénommerons l'**hypothèse de Continuum Temporel** : cette hypothèse suppose que les structures temporelles qui encodent l'identité d'une langue, les attitudes, les catégories lexicales ou des séquences de tons purs forment un continuum. Un corollaire direct de cette hypothèse est qu'il existe un seul et même système capable de traiter ces structures temporelles, le long de ce continuum. Les recherches précédentes de notre groupe ont développé un modèle d'apprentissage des séquences sensori-motrices, le réseau récurrent temporel (TRN). Ce modèle emploie un réseau d'unités dynamiques, avec une

distribution de différentes constantes de temps, reliées par des connexions récurrentes. La combinaison de ces principes autorise un système dynamique, qui est sensible en théorie à différentes échelles temporelles.

Notre premier chapitre présentera les travaux des neurosciences computationnelles pour le traitement des informations temporelles et inclura en détail le modèle utilisé (TRN). La problématique du traitement des informations inscrites dans le temps constitue ainsi le point de départ de notre travail. Le chapitre suivant exposera les travaux concernant la prosodie.

La partie expérimentale de cette thèse vérifiera donc que l'identification des langues, des attitudes, des catégories lexicales et de tons purs brefs peut être réalisée par un seul mécanisme, avec des hypothèses neurophysiologiques plausibles (TRN).

Trois tâches de traitement de parole et deux tâches de perception de tons purs évoqueront ainsi une organisation temporelle d'abord globale puis de plus en plus locale :

1. L'Identification Automatique des Langues (IAL) s'appuie sur la distribution globale de la prosodie d'une langue. Le système devra indiquer la langue parlée contenue dans un signal acoustique ;
2. L'Identification des Attitudes Prosodiques étudiera la prosodie définie pour des courtes phrases de 6 syllabes. Chacune de ces attitudes a été définie dans un cadre très strict en vue de leur synthèse (Morlec, Bailly et Aubergé, 2001). Le système devra indiquer l'attitude ou la modalité d'une phrase à partir du contour intonatif ;
3. La Catégorisation Lexicale portera sur la prosodie des mots. Le système devra distinguer les mots de fonction des mots de contenu ;
4. Ce dernier point permettra d'examiner les conséquences d'un Dysfonctionnement lors du Traitement Temporel Auditif. Ce dysfonctionnement peut être obtenu par une augmentation des valeurs des constantes de temps, caractérisant le réseau, simulant ainsi un trouble biologique des neurones.

Tout au long de la thèse, plusieurs points doivent être considérés : chacun des domaines abordés représente un domaine de recherches indépendant et fortement développé. L'objectif principal de cette thèse n'est pas de démontrer des résultats supérieurs à ceux obtenus par des approches plus spécifiques. Mais l'objectif est de prouver que le TRN peut obtenir des performances voisines de celles des êtres humains, et plus particulièrement d'auditeurs « naïfs » pour la parole testée, comme des nourrissons, en respectant la « contrainte temporelle » que nous allons maintenant décrire. Ainsi le TRN peut être la simulation d'une ressource cérébrale potentiellement utile pour résoudre certaines tâches intervenant lors de l'acquisition du langage.

Pour chacun de ces domaines, la durée temporelle est extraite, si bien que le signal analogique de la parole est souvent transformé en une séquence de symboles discrets (phonème, syllabes ou mots dont la durée est codée sous forme symbolique ou numérique). Ces symboles sont d'une part dépendants de la tâche à réaliser (syllabe pour le traitement du contour intonatif, mots pour l'apprentissage de la syntaxe, etc.…) et d'autre part définis à partir de connaissances précises de la langue étudiée, qui ne sont

pas directement accessibles à un nouveau-né. Ainsi, la contrainte temporelle nécessite que la structure temporelle du signal acoustique reste dans sa forme analogique à l'entrée du système, sans conversion des données en durée codée symboliquement. Cette contrainte reflète ainsi un traitement réaliste du temps.

Cette thèse apporte deux contributions : premièrement, en exploitant de façon « naïve » les données acoustiques pour catégoriser les structures prosodiques, une même architecture, le modèle TRN, modélise certaines des tâches réalisées par le nouveau-né, confronté à la prosodie comme point d'entrée perceptuelle de sa langue. Effectivement, le TRN reste compétitif avec des approches alternatives, proposées dans chacun de ces domaines. Deuxièmement, des indices temporels sont identifiés et validés dans deux langues pour distinguer les catégories lexicales.

Mais ne laissons pas le temps filer plus longtemps…

Chapitre Un Le Traitement du Temps

« Le Temps est le sens de la vie » Paul Claudel

comme on dit le sens d'un cours d'eau, le sens de l'ouïe, … et le sens de cette thèse, ou tout du moins son origine.

I.Les défis posés par la dimension temporelle

Toutes les activités humaines ainsi que les modalités sensorielles sont intimement liées à la dimension du temps. Le système nerveux traite toutes les informations sensorielles en intégrant leur composante temporelle. Des données expérimentales et théoriques suggèrent que le temps intervient même lors d'événements a priori statiques (Richmond, Otpican et Spitzer, 1990 ; McClurkin, Optican, Richmondet Gawne, 1991 ; Middlebrooks, Clock, Xu et Green, 1994 ; Mechler, Victor, Purpura, et Shapley, 1998 ;Prut, Vaadia, Berman, Haalman, Solvin,. et Abeles, 1998 ; Buonomano et Merzenich, 1999). L'ouïe est particulièrement tributaire du temps où des événements complexes comme la parole ou la musique doivent suivre une structure séquentielle. Étant donné la pression engendrée par le temps sur notre environnement, l'informatique a naturellement été conduite à le prendre en compte dans ses modèles.

Nous résumerons d'abord la représentation du temps dans les modèles informatiques. Ensuite, nous nous intéresserons aux réseaux de neurones, ceux traitant

les structures séquentielles, puis ceux capables de prendre en compte la durée des évènements. Enfin, nous décrirons un modèle apparteant à cette seconde catégorie : le réseau récurrent temporel TRN employé dans cette thèse.

I.1.Les modèles de représentation du temps

La représentation du temps est un des problèmes les plus cruciaux pour tout système informatique qui tend à décrire le monde. Les bases de données, les simulations, les systèmes experts, et les applications de l'Intelligence Artificielle en général peuvent se retrouver confronter à ce problème. Aussi, l'informatique a proposé de nombreuses solutions pour exprimer le temps dans ses programmes (Allen, 1991).

La représentation absolue du temps, sous forme d'horloge numérique est la forme la plus souple du point de vue de l'algorithmique traditionnelle. Ainsi les chiffres (1990 110 10 4 50) représentent le 110éme jour de l'année 1990, à 10 : 04et 50 secondes. Cependant, cette première représentation ne donne pas accès aux durées.

L'Intelligence Artificielle a alors employé des graphes dont les arcs indiquent les relations temporelles entre différents événements, les nœuds du graphe. Ces graphes peuvent représenter les durées, tout en maintenant un ordre partiel des événements. Chaque point du graphe dispose de deux coordonnées : le début et la fin d'un événement.

Jusqu'à présent le temps seul était pris en compte, mais il est important d'associer des objets et leur état (prédicat) à un instant possible. Vendler (1957) avait noté une distinction importante entre deux types de prédicats. Effectivement, il est possible que si un prédicat s'applique à un intervalle de temps donné, soit il s'applique à tous les temps compris dans cet intervalle (cas homogène), soit aucun des temps n'est caractérisé par le prédicat. Ainsi, si un objet est vert pendant une période, il sera vert pour chaque instant de cette période. En revanche, si le soleil se lève pendant un intervalle donné, il sera immobile à chaque instant.

La philosophie et la linguistique (par exemple, tense logic, Prior (1967)) appuyé par l'informatique théorique (logique dynamique, Pratt, 1978 ; Harel, 1979) ont proposé une nouvelle représentation du temps : des prédicats sont utilisés pour caractériser les relations temporelles entre les événements. Par exemple (passé(verte(grenouille))) signifie qu'il y a eu un temps où une grenouille a été verte. Cette représentation logique a été utile pour le langage, ou pour la sémantique formelle des programmes. Cependant elle ne permet pas non plus de capturer des situations complexes, telles que la persistance des événements.

En toute logique, si l'on a une connaissance sur un objet à un moment donné, il n'est pas possible d'en tirer des conclusions sur son état dans le futur. Cependant, la réalité est tout autre. Si l'on gare sa voiture à un moment donné, il est hors de question de vérifier à chaque instant la présence de celle-ci sur son lieu de garage. En outre, si la police appelle pour dire qu'elle a retrouvé cette voiture quelque part ailleurs, le prédicat doit pouvoir être modifié. Une technique tenant compte de telles assomptions doit être non-monotone (Dean et Kanazawa, 1988 ; Weber, 1989).

Tenir compte du temps dans un système de traitement de l'information pose deux

grandes contraintes. Tout d'abord, ce système doit pouvoir gérer la succession des différents événements qui doivent être traités de manière séquentielle — il est alors question de traitement séquentiel. Ensuite, la prise en compte des durées par le système doit pouvoir se greffer sur ce traitement séquentiel. Ainsi, si la durée des événements est pertinente pour la tâche à executer, le système doit être en mesure de traiter la structure temporelle.

Quelles sont les solutions proposées par les réseaux de neurones pour traiter :

1. Le caractère séquentiel de l'information ?
2. Les durées des événements ?

I.2.Les réseaux de neurones face au traitement du temps

Historiquement, les réseaux de neurones ont été utilisés pour des problèmes de classification ou d'approximation de fonction sur des données statiques : la sortie d'un réseau ne dépend que de l'entrée actuelle. Pour pouvoir atteindre un rendu plus réaliste des réseaux de neurones biologiques, il faut prendre en compte l'aspect dynamique du problème : à un instant donné, la sortie dépend de l'entrée à l'instant courant, ainsi que des entrées et des sorties aux instants précédents, jusqu'à un certain ordre. Les informaticiens ont été amenés à étudier les modèles de traitement de séquences **pour des applications telles que l'identification et le contrôle de système dynamique,** la reconnaissance de motifs syntaxiques, l'induction de la grammaire (Kremer, 2001).

Les réseaux connexionnistes spatio-temporels peuvent se classer suivant la représentation du temps utilisée par ces modèles de traitement de séquence [3] . Cette représentation peut être :

- **Discrète :** le temps est réduit à une succession d'évènements symbolique (mots, syllabes, etc.). Ces modèles considèrent que chaque événement a la même durée. Elle n'est pas donc prise en compte par ces modèles, qui proposent uniquement un traitement séquentiel.
- **Temporelle** : cas des modèles biologiques actuellement étudiés qui s'appuient sur un échantillonnage très fin du temps [4] , sans tenir compte d'unité temporelle. Cette particularité permet aux modèles de tenir compte de l'organisation temporelle des données. Nous évoquerons dans ce cas l'expression de traitement réaliste du temps (ou contrainte temporelle), dans la mesure où elle permet de s'approcher au mieux de notre perception physiologique du temps.

[3] Horne et Giles (1995) ont développé une taxonomie pour comparer les réseaux connexionnistes spatiaux temporels à partir de leur architecture. Ils isolent les réseaux utilisant des unités cachées pour coder l'état du réseau. Mozer (1994) et Kremer (2001) proposent une classification plus complète de ce type de réseaux.

[4] De l'ordre de quelques millisecondes.

II.Traitement de séquences discrètes

Les modèles de traitement de séquences discrètes se distinguent en deux catégories :

1. Les modèles théoriques fournis par l'informatique, et dont le but est de pouvoir apprendre des séquences discrètes ;
2. Les modèles biologiques qui tiennent compte des contraintes biologiques comme l'anatomie ou la physiologie, inpirées par les études en neurosciences.

II.1.Modèles théoriques de traitement de séquences discrètes

Deux types de représentations du contexte peuvent être envisagés pour les modèles théoriques :

- De façon **externe** au modèle : Une sorte de "spatialisation" du temps est effectuée (Time-Delay Neural Networks ; TDNN). Les séquences temporelles d'événements sont transformées en séquences spatiales. Seulement, il n'est pas clair que les systèmes biologiques puissent utiliser de tels registres.
- De façon **interne** au modèle : Les évènements passés sont encodés à l'aide de connexions supplémentaires (connexions récurrentes).

Les modèles biologiques intègrent dans leur architecture des connexions récurrentes afin de traiter le contexte de manière interne au modèle.

II.1.1.Représentation du temps externe au modèle

Le temps est représenté par une métaphore spatiale. Les entrées du réseau sont mémorisées dans un registre particulier (Window In Time Memory [5]). Le premier événement temporel correspond à la première coordonnée du vecteur, et ainsi de suite. Ainsi si le symbole d'entrée à un instant *t* est *X(t)*, alors la fenêtre correspondant à 5 pas sera constituée des 5 symboles séquentiels :*(X(t), X(t-1), X(t-2), X(t-3), X(t-4))*. La dimension du vecteur des données d'entrées est donc fonction de l'ordre sériel des données à analyser.

La réussite de ce réseau dépend de l'encodage symbolique retenue pour figurer le temps. NETtalk (Sejnowski et Rosenberg, 1986) constitue la première application d'un réseau dynamique, qui apprend à "prononcer" l'anglais.

Cette approche a également été testée avec deux fenêtres temporelles (une pour les entrées et une pour les sorties : Nonlinear AutoRegressive with eXogenous inputs [6] : Lin, Horne, Tiño et Giles, 1996). Après avoir retenu l'information des couches d'entrées et de

[5] Ces réseaux sont équivalents à des machines finies (Kohavi, 1978, cité dans Giles and Horne, 1994).

sortie, un troisième type de réseau mémorise les activations des couches cachées. Un réseau de deux couches a été implémenté pour la reconnaissance de phonèmes (Time-Delay Neural Network memories : Waibel, 1988 ; Waibel, Hanazawa, Hinton, Shikano et Lang, 1989). Cette approche donne de très bons résultats, supérieurs à ceux obtenus avec des Modèles de Markov Cachés (Hérault et Jutten, 1994).

Afin de pouvoir palier au nombre fini d'activations enregistrables, l'historique est codé de manière incrémentale en suivant une décroissance soit exponentielle [7] (Feedforward Exponential Decay Memories) soit calquée sur une fonction Gamma (Gamma Memories).

II.1.2.Représentation du temps interne au modèle

Une des difficultés pour les réseaux connexionnistes est de représenter des structures de données récursives (liste ou arbre) dont la taille est variable. Les réseaux doivent être capables de produire des représentations distribuées compactes pour des structures composées, ainsi que des mécanismes pour extraire ces structures, à partir de ces représentations compactes.

Pour ce faire, les réseaux suivants mémorisent les informations passées (c'est à dire le contexte). Les couches cachées reçoivent l'activation émise par d'autres neurones à une étape antérieure. Ainsi, le temps est figuré par ces effets, plutôt que par une métaphore spatiale. Les connexions qui permettent ces transmissions sont appelées connexions récurrentes.

Cette activation peut provenir des neurones de sortie (Jordan, 1986b ; Jordan, 1986a ; Jordan, 1990), ou à la fois des entrées courantes, et des activations précédentes du réseau (copiées dans une couche dite couche de contexte). Almeida (1987) et Pearlmutter (1988) ont développé des méthodes d'apprentissage pour les réseaux récurrents (recurrent back propagation).

De nombreux modèles ont repris ce principe [8] : Real Time Recurrent Learning Network (Williams et Zipser, 1989), Simple Recurrent Network (Elman, 1990), Recurrent Cascade Correlation (Fahlman, 1991), Recurrent Auto-Associative Memory (Pollack, 1989 ; 1990 ; 1994 ; Voegtlin, 2002b), Auto-Associative Recurrent Network (Maskara et Noetzel, 1992).

Un neurone peut également être influencé par deux autres neurones : l'activation de l'une des unités est modulée par une seconde **unité (** Second-Order Context Memory : Rumelhart, Hinton et Williams, 1986). Afin d'apprendre **un nombre plus important** d'associations spatio-temporelles, d'autres techniques d'apprentissage ont été développées (Long Short-term Memory : Hochreiter et Schmidhuber, 1997 a&b), ainsi que dans le cas d'apprentissage non supervisé (Voegtlin, 2002a&b).

[6] Les réseaux NARX sont équivalents aux machines de Turing (Siegelmann , horne et Giles., 1997).

[7] Ce modèle a été utilisé pour la prédiction des signaux de paroles (Poddar et Unnikrishnan, 1991).

[8] Ces réseaux sont équivalent à un automate à état fini (Kremer, 1995).

II.1.3.Critique des modèles théoriques

Une limitation importante de la descente du gradient a été identifiée indépendamment par Hochreiter (1991) et Bengio, Simard et Frasconi (1994). Lorsque la durée des séquences temporelles augmente, l'influence des composantes initiales de la séquence a de moins en moins d'impact sur la sortie du réseau. Cela a conduit à la définition du gradient partiel qui définit le changement des poids de plus en plus proches de zéro, au fur et à mesure que la séquence apprise augmente.

Bien que ces modèles puissent reproduire des données comportementales (Cleeremans et McClelland, 1991 ; Elman, 1990), ils ne respectent pas des contraintes anatomiques ou physiologiques. Quelles sont les structures cérébrales impliquées dans le traitement de séquences discrètes ?

II.2.Inspiration et contexte neurologique

Le cortex préfrontal joue un rôle majeur pour l'analyse et le contrôle des informations ordonnées dans une structure sérielle. Effectivement, les sujets avec des lésions frontales ont montré des performances atténuées dans des tâches séquentielles de pointage, de reconnaissance de l'ordre sériel, ou de jugement sur l'ordre d'apparition (Beiser et Houk, 1998).

Les patients atteints de la maladie de Parkinson et ceux touchés par le syndrome de Huntington éprouvent des difficultés pour le contrôle moteur, mais aussi lors de tâches cognitives. Certains de ces déficits sont similaires aux déficits d'ordonnancement observés chez des patients fronto-lésés (Beiser et Houk, 1998).

Les singes ayant subi des lésions bilatérales des aires 46 et 9 ont des difficultés pour la maîtrise séquentielle de nouveaux stimuli (Petrides, 1991). Les enregistrements des cellules dans le primate en train d'exécuter des tâches de reproduction de séquences montrent l'importance des ganglions de la base et du cortex préfrontal pour le traitement sériel.

Les neurones des régions préfrontales, et proches de l'aire oculomotrice frontale (FEF) et du noyau caudé sont sensibles à l'ordre sériel des séquences apprises (Barone et Joseph, 1989 ; Funahashi, Inoue et Kubota, 1993 ; Kermadi et Joseph, 1995 ; Kermadi, Jurquet, Arzi et Joseph, 1993).Les réponses qui sont générées par les séquences avant la réponse motrice peuvent représenter la conversion des séquences temporelles des entrées sensorielles en un motif spatial d'activité neuronale. De la même façon, des unités de l'aire oculomotrice frontale, noyau caudé, et du pallidum ont trait à l'ordre sériel des séquences motrices (Barone et Joseph, 1989 ; Kermadi et Joseph, 1995 ; Kermadi et coll., 1993). Cette activité peut représenter les commandes de la conversion inverse, c'est-à-dire d'un motif spatial vers un motif temporel, qui sera à l'origine du mouvement. Ensemble, ces études apportent des preuves de l'existence d'un mécanisme effectuant ces conversions, donnant ainsi crédit aux idées qu'avaient exprimées Lashley en 1951.

Les réponses observées dans le cortex préfrontal des primates montrent que celui-ci pourrait être équivalent à une mémoire de travail spatiale (Funahashi, Bruce et

Goldman-Rakic, 1989 ; 1990 ; Fuster et Alexander, 1971 ; Goldman-Rakic, 1995 ; Petrides, 1991). Les boucles cortico-thalamiques peuvent soutenir des activations semblables à une mémoire de travail (Dominey et Arbib, 1992 ; Hikosaka, 1989 ; Houk et Wise, 1995). Les boucles cortex – thalamus sont seulement une des possibilités pour assurer une activité soutenue dans la mémoire de travail. Au moins quatre boucles de retour vers le préfrontal sont envisageables : cortico-corticale avec le pariétal postérieur, cortico-corticale à l'intérieur du cortex préfrontal, cortex – cervelet vers les noyaux dentés et des boucles à l'intérieur du striatum vers les ganglions de la base (Houk, 1997). Le cortex préfrontal peut garder en mémoire des informations de l'ordre d'une seconde (Fuster, 1993), tandis que la mémoire locale des ganglions de la base serait de l'ordre de 10 ms à 100 ms. Une forme locale de mémoire de travail est représentée par des connexions réciproques entre le pallidum externe et le noyau subthalamique.

Les études d'imageries fonctionnelles (IRMf) ont également prouvé que le cortex préfrontal humain faisait preuve d'une activité similaire en tant que mémoire de travail (Fiez et coll., 1996 ; Jonides et coll., 1993 ; McCarthy et coll., 1994).

Les décharges émises pendant la période de rétention ont aussi été identifiées dans le noyau caudé (Hikosaka, Sakamoto et Sadanari, 1989 ; Schultz et Romo, 1992), la substance noire (SNr : Hikosaka et Wurtz, 1983) et le thalamus (Fuster et Alexander, 1971). Il a été suggéré que le mécanisme permettant l'intégration temporelle puisse être considéré comme une extension de celui de la mémoire de travail (Fuster, 1985 ; Goldman-Rakic, 1987).

Quels sont les modèles inspirés par ces connaissances des structures cérébrales ?

II.3.Modèles biologiques de traitement de séquences discrètes

Récemment des modèles du traitement de l'information dans les ganglions de la base ont été créés : pour le traitement sériel, la sélection de l'action, et l'apprentissage par renforcement. Pour les modèles de traitement sériel, les ganglions de la base jouent un rôle central pour la génération de motifs d'activités (Berns et Sejnowski, 1998). Ces modèles n'ont pas recours aux mécanismes d'apprentissage par rétropropagation (Cleeremans et McClelland, 1991).

De nombreuses simulations informatiques de substrats neuronaux ont évoqué les fonctions sensorimotrices des ganglions de la base (*cf.* Gillies et Arbuthnott, 2000) dans le cas du modèle de Beiser et Houk (1998) et dans le contexte de l'algorithme d'apprentissage par Différence Temporelle (Temporal Difference Learning) appliqué dans des modèles Acteur-Critique (Joel et coll., 2002 ; Suri et Schultz., 1998).

II.3.1.Le modèle de Beiser et Houk (1998)

Beiser et Houk (1998) proposent un réseau de neurones inspiré de circuits spécifiques reliant le cortex préfrontal aux ganglions de la base, deux structures impliquées dans les traitements sériels. Ce réseau est la réalisation du modèle conceptuel de Houk et Wise (1995). Ce modèle a été testé sur une tâche de reproduction de séquences. Il est capable d'effectuer la transformation des entrées sensorielles en motifs spatiaux (encodage). La

transformation inverse (décodage) n'a pas été modélisée. Lors du test du modèle, ces réponses sont comparables à des enregistrements de neurones du cortex préfrontal et des ganglions de la base [9] , durant une tâche de reproduction de séquences.

La simulation du cortex préfrontal utilise des neurones qui représentent les évènements dans la couche d'entrée et des neurones récurrents. Chaque stimulus est représenté par un signal qui a une valeur unitaire lors de la présence du stimulus, et une valeur nulle pour tous les autres cas [10] . Les neurones codant les évènements ressemblent aux neurones de fixation visuelle, dans le sens où leurs décharges sont soutenues tant que le stimulus reste dans le champ visuel (Goldberg et Colby, 1989). La couche représentant le noyau caudé forme un motif spatial tenant compte des informations du signal actuel et passé (couche de contexte).

Pour démontrer les propriétés algorithmiques du modèle, un groupe de six séquences de trois éléments A, B, et C (i.e., ABC, ACB, BAC, BCA, CAB, et CBA) a été présenté au réseau. Ces six séquences forment quinze séquences contextuelles de longueur variable (A, B, C, AB, AC, BA, BC, CA, CB, ABC, ACB, BAC, BCA, CAB, and CBA). Quinze motifs distincts sont issus du réseau pour les quinze séquences contextuelles, ce qui représente un codage non ambigu de l'information sérielle des entrées. L'encodage en motifs spatiaux ne requiert pas d'apprentissage adaptatif, bien qu'il puisse améliorer les performances. Cependant, certaines distributions des poids des connexions ne permettent pas de produire des motifs distincts (par exemple, toutes les unités sont saturées) [11] .

En outre, les champs récepteurs correspondent à ceux rencontrés dans les études d'enregistrement de neurones isolés. Les auteurs proposent une étude qualitative des champs récepteurs des neurones de la couche préfrontale. Ainsi, l'unité 2 répond à un événement C, mais également à toutes les séquences commençant par C. Quelques unités répondent à un seul stimulus donné de taille 1, 2 ou 3. Trois types de réponses constatées dans le modèle ont été effectivement observés en neurophysiologie :

1. des neurones réagissant à un seul élément ont été enregistrés dans le cortex préfrontal par Funahashi et coll. (1993) ;
2. des neurones répondant à des stimuli de rang 2 ou 3 dans une séquence appartiennent aux régions de l'aire oculomotrice frontale (Barone et Joseph, 1989) et du noyau caudé (Kermadi et Joseph, 1995 ; Kermadi et coll., 1993).
3. La plupart des unités réagissent à plusieurs séquences (cortex préfrontal, Funahashi et coll., 1989, 1993).

[9] Une séquence donnée par des ampoules lumineuses doit être reproduites par un primate non-humain après un court instant.

[10] Des représentations similaires des stimuli sont apparues dans d'autres études : 'complete serial compound stimulus' (Sutton & Barto, 1990) or 'spectral timing mechanism' (Brown, Bullock et Grossberg, 1999).

[11] Une augmentation des constantes de temps de 15 ms à 150 ms a permis d'améliorer les performances de 270 à 460 réseaux parfaits (i.e. produisant un motif distinct pour chacune des 15 séquences). Un mécanisme supplémentaire pourrait être considéré pour guider le choix de ces constantes (Beiser et Houk, 1998).

La boucle de la mémoire de travail reste à un niveau constant pendant la présentation d'un événement d'une séquence, ce qui correspond à des données neurophysiologiques (Funahashi et coll., 1989 ; Fuster et Alexander 1971). Cette activité pourrait créer des encodages relativement insensibles à la vitesse de présentation des indices, encodant ainsi d'abord la structure sérielle des séquences.

II.3.2.Apprentissage par Différence Temporelle (TD)

L'algorithme d'apprentissage par Différence Temporelle est un apprentissage par renforcement [12] (Suri, 2002). Son développement a été fortement influencé par les études d'apprentissage chez les animaux (Sutton et Barto, 1998). Il utilise un mécanisme d'estimation temporelle, qui prédit le temps d'arrivée d'une récompense lors de paradigmes pavloviens.

Le signal d'erreur du modèle TD était purement hypothétique, jusqu'à ce qu'il soit découvert que l'activité des neurones dopaminergiques de la substance noire et des aires ventrales et tegmentales ressemble au signal de récompense de la prédiction (Suri, 2002 ; Suri et Schultz, 1999). Les activations des neurones dopaminergiques ressemblent à celles décrites par les modèles d'apprentissage fondés sur les différences temporelles (Suri, 2002).

Des troubles observés dans la transmission de la dopamine perturbent le mouvement sériel chez les sujets humains (Suri, 2002). Etant donné les capacités d'apprentissage pour les séquences d'action de l'algorithme TD, cette dernière observation permet de conclure que l'activité des neurones dopaminergiques sert de signal prédictible pour l'apprentissage, dans une architecture biologique (Sutton et Barto, 1998).

L'apprentissage dépend du degré d'imprédictibilité des récompenses (Suri, 2002). Seules les récompenses apparaissant de façon imprévisibles vont renforcer l'apprentissage. La courbe d'apprentissage suit une asymptote, quand toutes les récompenses sont prévisibles. L'erreur ou la différence entre l'apparition de la récompense et sa prédiction entre en jeu lors de l'apprentissage de nombreux modèles de réseaux de neurones (Sutton et Barto, 1998). Suri, Bargas et Arbib (2001) proposent un modèle capable d'apprendre des séquences sensori-motrices, avec en plus des capacités de planification : ils forment de nouvelles chaînes associatives, et choisit son action en fonction des sorties prédites par ces chaînes associatives.

II.3.3.Système Acteur-Critique

L'algorithme d'apprentissage décrit précédemment est principalement utilisé dans le cadre des modèles Acteur-Critique. Un sous-réseau dénommé Acteur apprend des actions de manière à maximiser la somme pondérée des futures récompenses, qui est calculée à chaque itération par un autre sous-réseau, la Critique (Barto, 1995). La Critique est adaptative dans le sens où elle prédit les récompenses à partir des entrées courantes et de l'activité de l'acteur, en comparant la prédiction avec les récompenses réelles. L'erreur entre deux prédictions est appelée erreur de différence temporelle. Elle est

[12] Il converge vers une solution optimale (Dayan & Sejnowski, 1994).

utilisée pour mettre à jour les poids des connexions de la Critique (Sutton, 1988)

Niv, Joel, Meilijson et Ruppin (2002) ont appliqué des algorithmes génétiques au système Acteur-Critique pour modéliser la prise de décision lors de la récolte du nectar chez les abeilles. Les troubles des patients parkinsoniens peuvent également être modélisés, en diminuant la valeur de la vitesse d'apprentissage, ce qui simule une diminution de la dopamine (Berns et Sejnowski, 1998).

II.4.Conclusion

Dans le modèle de Beiser et Houk (1998), les codes neuraux ne tiennent compte que de la dimension spatiale. Dans celui développé par Suri et Schultz (1999), le temps est représenté de façon discrète. Les réseaux ne reçoivent pas les informations au cours du temps, mais à des instants précis dépendant de la tâche à executer. La méthode TD semble être bien adaptée à l'apprentissage de séquences abstraites (Berns et Sejnowski, 1998 ; Beiser et Houk, 1998), mais son application à l'étude des structures temporelles, en particulier pour des structures réelles, reste à démontrer. Est-il possible d'adapter ces modèles pour qu'ils puissent tenir compte de la durée des intervalles ? Quels sont donc les modèles, qui ont été démontrés comme étant sensibles à la structure temporelle ? Sur quelles données physiologiques et comportementales se basent-ils ?

III.Traitement de séquences temporelles

Pour ces modèles, le temps est envisagé comme un degré de liberté supplémentaire dans les représentations neurales. Les algorithmes employés par les modèles théoriques ont été modifiés pour pouvoir tenir compte des variations de durées des éléments transmis au réseau. Deux implémentations ont été proposées (Pearlmutter, 1995) :

1. Traiter les connexions comme un ensemble nouveau à chaque itération (Back Propagation Through Time) ;
2. Conserver la contribution de chaque connexion pendant chaque itération (Real Time Recurrent Learning).

Cependant, ces méthodes ne concordent pas avec les données neurophysiologiques, dans la mesure où les moyens de traitement sont limités. Quelles solutions ont été procurées par les études en neurophysiologie pour répondre à cette question du traitement des séquences temporelles ?

Nous présenterons d'abord les connaissances disponibles sur le traitement du temps dans les structures cérébrales, avant de décrire quelques modélisations inspirées par ces études. Enfin, nous terminerons avec une présentation détaillé du modèle de réseau récurrent temporel, utilisé dans cette étude : le TRN.

III.1.Inspiration et contexte neurologique

Un certain nombre de question ont été posées lors du traitement du temps par les mécanismes cérébraux. Celles que nous allons décrire concernent les échelles de temps, le type de codage, les composants de ce système et plus particulièrement si il pourrait s'agir d'un mécanisme central ou distribué.

III.1.1.Echelle temporelle

L'échelle temporelle de l'information traitée par le système nerveux varie suivant plusieurs magnitudes : de quelques microsecondes et millisecondes, jusqu'à plusieurs secondes (Buonomano, 2000). Les durées de l'ordre de quelques millisecondes concernent une bonne partie du traitement sensoriel [13] : le système cérébral peut discriminer des durées différentes pour des tons purs (Wright et coll., 1997 ; Wright, Buonomano, Mahncke et Merzenich, 1997) et pour de la parole (Tallal, 1994 ; Shannon et coll., 1995).

En outre, des données expérimentales ont montré que certains neurones sensoriels répondent sélectivement à des motifs temporels entre 10 et 100 ms (Buonomano, 2000). Chez le singe, il s'agit de neurones sensibles aux appels (Rauschecker, Tian et Hauser,1995 ; Wang, Merzenich, Beitel, et Schreiner, 1995), aux intervalles et aux durées (Riquimaroux, 1994 ; He, Hashikawa, Ojima et Kinouchi, 1997), chez les oiseaux, des neurones sensibles aux chants (Margoliash, 1983 ; Lewicki et Arthur,1996), et chez l'humain, des neurones sensibles aux mots (Creutzfeldt, Ojemann et Lettich, 1989).

III.1.2.Type de codage

L'encodage de l'information dans les flux des impulsions nerveuses n'est pas non plus totalement éclairci. Deux hypothèses prédominent pour expliquer le codage de l'information dans la distribution des impulsions nerveuses (Doerksen, 2000 ; Liaw et Berger, 1996) :

1. le codage par fréquence privilégie la fréquence moyenne d'émission de décharges d'un neurone pour coder l'information.
2. le codage temporel emploie la structure temporelle des impulsions pour porter l'information. Il permet de coder beaucoup plus d'informations, mais il est ainsi beaucoup plus sensible au bruit. Des observations biophysiques ont montré que le flux des décharges est bruité, mais il reste possible que ce bruit contienne des informations significatives (Doerksen, 2000).

Etant donné l'immense variabilité observée lors de l'émission des impulsions, le codage basé sur la vitesse est le plus souvent retenu [14] .

[13] Certains codes temporels intrinsèques ont cette même échelle de durée (Mechler, Victor, Purpura et Shapley., 1998).

[14] Tout le système emploie alors principalement des motifs spatiaux.

III.1.3.Description du système de traitement de séquences temporelles

Jusqu'à présent, peu de choses sont connues sur les mécanismes neuronaux qui sont sensibles aux durées de l'ordre de la millisecondes (Ivry, 1996 ; Gibbon, Malapani, Dale et Gallistel, 1997) et ce, même pour de simples discriminations de durées.

Le cervelet et les ganglions de la base interviennent dans une large variété de tâches temporelles. L'architecture du cervelet pourrait être adaptée au calcul précis des relations temporelles entre différentes entrées, et entre différents motifs d'entrée et de sortie (Hazeltine, Grafton et Ivry, 1997). Des mécanismes pré- et post-synaptiques relativement lents permettent aux neurones de construire des représentations de l'information temporelle. Ainsi, les neurones de l'hippocampe représentent et traitent l'information au moyen de motifs d'activations spatio-temporels (Liaw et Berger, 1998).

De nombreux mécanismes ont été proposés pour traiter le temps (Buonomano, 2000) :

- Des délais temporels (delay lines : Braitenberg, 1967 ; Tank et Hopfield, 1987) ;
- Des oscillateurs (Miall, 1989 ; Ahissar et coll., 1997) ;
- Des réseaux dynamiques (Buonomano et Mauk, 1994) ;
- La plasticité des synapses à court terme (Buonomano et Merzenich, 1995) ;
- Une horloge interne commune [15] (Treisman,1963).

III.1.4.Un système centralisé ou distribué ?

Il est toujours difficile de déterminer si le système qui permet de traiter les durées de 10 ms à 100 ms est central ou distribué dans différentes régions du cerveau (Karmarkar et Buonomano, 1998). Une des difficultés dans le traitement du temps concerne le traitement des paramètres non temporels comme, par exemple, la hauteur d'un ton. Ainsi la circuiterie neurale est-elle la même pour toutes les fréquences, ou est-elle spécifique de chaque fréquence ? Dans le premier cas, le système est central, dans le second cas, il s'agit d'un système distribué. Contrairement aux modèles centraux, les modèles distribués impliquent que des systèmes différents traitent l'information temporelle en fonction de la modalité demandée.

De la même manière, on peut imaginer qu'un système neural de « chronométrie » pourrait être spécifique de la durée qu'il a mesurée.

Des études sur des patients avec des lésions dans le cerebellum (Ivry et Keele, 1989), le cortex pariétal (Harringtonet coll., 1998a), et les ganglions de la base (Harrington et coll.,1998b) ont tous montré des déficits de traitement temporel, souvent pour des tâches perceptuelles et motrices. Ces études sont alors généralement interprétées comme des preuves d'un mécanisme centralisé.

Une autre façon de répondre à ces questions est d'étudier les processus d'apprentissage pour une tâche auditive de discrimination d'intervalles (Nagarajan et coll., 1998). Ainsi les sujets doivent apprendre à distinguer deux stimuli contenant deux tons

[15] Les problèmes temporels accéderaient à la même horloge interne quelle que soit la modalité à laquelle ils sont liés.

séparés par un intervalle. La seule différence entre ces deux stimuli tient dans la durée de l'intervalle entre les deux tons. Un entraînement de plusieurs jours permet d'identifier l'ordre dans lequel sont présentés ces stimuli, la différence de durée entre les deux intervalles se réduisant petit à petit.

Des études psychophysiques sur la discrimination d'intervalles ont exhibées des généralisations d'apprentissage entre deux modalités (Nagarajan et coll., 1998 ; Westheimer 1999) ou deux canaux (Wright et coll., 1997b). Cet entraînement se généralise suivant les différentes fréquences, mais pas pour des durées d'intervalles différentes (Wright et coll., 1997b). Il est même possible de transférer l'apprentissage vers des tâches sensori-motrices (Meegan et coll., 2000).

Les apprentissage observés dans ces études peuvent être interprétés de deux façons distinctes : soit la manière de mesurer le temps s'est améliorée, soit les capacités pour stocker et comparer les deux stimuli se sont améliorées. Nagarajan et coll. (1998) montrent qu'il s'agirait d'un système central, mais dédié à des intervalles spécifiques [16]. Des similarités marquées dans les caractéristiques temporelles, dans des tâches de production et de perception impliquent un système temporel commun. Les calculs dédiés au temps doivent être distribués dans le cerveau, mais des remarques récentes suggèrent des rôles spécifiques à différentes structures neurales.

III.2.Modèles neuromimétiques pour le traitement des séquences temporelles

L'avancement des recherches en neurophysiologie permet le développement de nouveaux modèles neuromimétiques, notamment en ce qui concerne le traitement implicite de l'information temporelle. Parmi ces modèles, on peut citer la colonne de mémorisation à court terme, la triade synaptique (Dehaene et Changeux, 1989), le modèle de propagation guidée (Béroule, 1985) et la colonne corticale (Burnod, 1988 ; Alexandre, Guyot, Haton et Burnod, 1991). Le modèle de colonne corticale a servi de base à la conception d'une carte neuronale (TOM, Temporal Organization Map) qui a été testée avec succès en reconnaissance de la parole (Durand, 1995). Les réseaux que nous allons décrire traduisent une activité temporelle en activité spatiale, soit en tenant compte de connexions dont la valeur reste constante au cours du temps, soit en adoptant des connexions dynamiques.

III.2.1.Connexions constantes au cours du temps

Les tous premiers modèles de la détection d'intervalle étaient constitués d'un retard temporel, établi sur la conduction axonale. Cependant, même si l'on suppose que les fibres parallèles du cerebellum agissent comme des délais (Braitenberg, 1967), il n'existe pas de données expérimentales montrant des délais de quelques millisecondes. La plupart des modèles fondés sur des retards ne peuvent pas discriminer des séquences,

[16] L'apprentissage présenté dans cette étude se généralise depuis la durée d'un silence (un intervalle) vers la durée d'un événement sonore.

alors que cette capacité permettrait de distinguer des stimuli complexes comme des vocalisations animales ou de la parole (Buonomano, 2000).

Buonomano et Merzenich (1995) ont proposé que des propriétés neuronales dépendantes du temps soient à la base du traitement temporel. Une modification des poids d'un circuit, composé de deux synapses, permet d'ajuster le réseau à un intervalle, sans qu'une constante de temps soit changée. Une estimation temporelle peut donc être dérivée d'un réseau de neurones formels, sans que celui-ci ait recours à différentes constantes de temps (Buonomano et Merzenich, 1995 ; 1999).

La règle de Hebb ne facilite pas la généralisation d'un réseau continu dans le temps, car celui-ci change continuellement d'état après avoir répondu à un stimulus initial. La plasticité à long terme des synapses peut être à l'origine de la formation de champs récepteurs non seulement spatiaux, mais aussi temporels. En étudiant un réseau composé d'une seule couche, mais avec de nombreux neurones, Buonomano et Merzenich (1995) ont montré qu'il était possible de distinguer un grand nombre de stimuli temporels. Les poids de ce réseau étaient tirés aléatoirement, cependant, la distribution temporelle des unités était suffisamment étendue pour former un codage robuste, pour un nombre important de durées et de séquences temporelles. Leur réseau a été testé sur trois tâches : discrimination de fréquences, de séquences temporelles aléatoires, et de phonèmes. Il est sensible à l'ordre, aux intervalles et permet la discrimination de séquences simples.

III.2.2.Synapses dynamiques

Deux postulats sont communément acceptés par les modèles connexionnistes :

1. Les synapses sont représentées de manières statiques, pendant le potentiel d'action d'un neurone ;
2. Chaque neurone transmet le même signal à tous les neurones auxquels il est connecté.

Or, le système nerveux utilise des synapses dynamiques, qui transmettent un signal différent à chaque neurone connecté. Ces types de synapses se retrouvent dans les aires du cerveau dédiées à la reconnaissance auditive (Baker et Rao, Soumis).

L'utilisation des synapses dynamiques permet d'augmenter les capacités de calculs et de diminuer le nombre de neurones. Ces réseaux doivent discriminer des stimuli consistant en deux décharges, séparées par un intervalle variable. Ils rivalisent avec les réseaux de neurones artificiels, mis en avant par Wang et Alkon (1995) dont l'architecture est fondée sur trois réseaux de neurones, ainsi que par Buonomano et Merzenich (1995) qui utilisent un nombre beaucoup plus important de neurones (500). Cependant, les neurones de ce réseau sont moins fiables, puisqu'ils peuvent répondre aussi bien pour 40 ms ou 50 ms. Mais, selon les auteurs, la plupart des applications réalistes n'ont pas besoin de comparer d'aussi petites variations. Le recours à des synapses dynamiques permet donc de diminuer le coût informatique (Doerksen, 2000). Elles permettent aussi de simuler des filtres spatiaux et temporels [17] .

Un réseau simple utilisant des synapses dynamiques est créé pour reconnaître des mots isolés noyés dans un bruit de fond, à partir du signal brut [18] (Liaw et Berger, 1996). Narmavar, Liaw, Berger (2001) ont ajouté un filtre basé sur les ondellettes. Un algorithme génétique permet d'optimiser les paramètres (comme les constantes du temps) du réseau et assure la convergence de l'apprentissage Hebbien. Les réseaux de synapses dynamiques ont aussi été combinés avec un modèle de cochléogramme, pour apprendre quatre tons purs dont la fréquence fondamentale varie au cours du temps [19] (Näger, Storck et Deco, 2002).

Le prochain réseau que nous allons étudier ne contient pas de synapses dynamiques, et peut donc être comparé aux réseaux de la première catégorie.

III.3.Un modèle de réseau récurrent temporel (TRN)

Cette partie décrit le fonctionnement du modèle de réseau récurrent temporel employé dans cette thèse. Ce modèle est initialement basé sur le système frontostriatal du primate. Ce système inclut des connexions cortico-corticales récurrentes et des connexions modifiables entre le cortex et le striatum. Toutes les connexions sont statiques. Il a été développé pour apprendre des ordres sensorimoteurs et pour simuler l'activité neurophysiologique pendant une tâche d'apprentissage de séquences accomplie par des primates non-humains (Dominey et coll., 1995). Ce modèle a initialement été créé pour expliquer l'activité des neurones du cortex préfrontal, qui encode à la fois la position spatiale et l'ordre sériel des stimuli.

Cette section comporte la description de l'architecture du modèle qui permet d'encoder des séquences spatio-temporel en motifs spatiaux, la façon dont l'apprentissage est réalisé et un descriptif des différences par rapport aux réseaux récurrents classiques, comme celui d'Elman (1990).

III.3.1.Architecture

Le réseau TRN emploie des unités, appelées intégrateurs à fuite, qui modélisent le taux de décharges des neurones. L'architecture du réseau reprend celle d'Elman (1990), pour pouvoir encoder les évènements passés dans la représentation courante. Enfin, contrairement aux études précédentes (Dominey et coll., 1995 ; Dominey, 1995 ; Dominey et Ramus, 2000), le réseau sera employé comme un système d'encodage. Dans ce contexte, l'apprentissage lui-même se déroule en dehors du réseau, mais il peut être réintégré dans le modèle par l'utilisation d'une mémoire associative.

[17] Natchschläger, Maass et Zador, (2000) ont montré que le recours à des synapses dynamiques permet d'obtenir une approximation d'un filtre temporel de façon plus efficace par rapport à un « time delay network » (Back et Tsoi, 1993), tout en réduisant le nombre de paramètres ajustables.

[18] Douze mots d'une syllabe prononcés par deux locuteurs.

[19] Les voyelles seules ne pouvaient être apprises, car elles étaient trop stationnaires pour le modèle. C'est aussi la raison pour laquelle les tons pur ont une fréquence fondamentale montante ou descendante.

III.3.1.1.Les intégrateurs à fuite

Le modèle de neurone réaliste le plus simple est l'intégrateur à fuite, ou « leaky integrator ». Il modélise le comportement d'un neurone par l'utilisation d'une équation différentielle temporelle. L'activité de la membrane s'obtient alors par une équation différentielle liant l'entrée et l'activité. La sortie du neurone est ensuite calculée par rapport à cette activité.

Il comprend une variable interne qui décrit l'état du neurone. Cette variable décrit le potentiel de membrane *m(t)* au niveau de la zone d'initiation de la décharge. L'évolution temporelle de cette variable est décrite par l'équation différentielle suivante :

$$\tau * \frac{\partial m}{\partial t} = -m(t) + \sum_i w_i X_i(t) + h$$

Avec *h* le niveau restant ; τ la constante de temps ; Xi(t) la décharge de la $i^{éme}$ entrée, et w_i le poids de la connexion correspondante.

Ce modèle permet ainsi de laisser une trace de l'activité d'un événement. De ce fait, les traces de deux événements générés à des instants différents peuvent coexister grâce à cette activité rémanente et interagir par apprentissage.

Un modèle simple de décharge neuronale a été proposé par Lapicque (1907), qui a relié le potentiel de membrane précédent à un seuil ; ainsi une décharge était générée chaque fois que le seuil était atteint. Hill (1936) utilise un couple de neurones à fuite, l'un décrivant le potentiel de membrane et l'autre le seuil dynamique de décharge. Un neurone à fuite ne calcule pas les décharges elles-même, mais définit la vitesse de décharge comme une mesure variant continuellement, caractérisant l'activité de la cellule. La vitesse de décharge est approximée par une fonction sigmoïde du potentiel de membrane, $M(t) = \sigma(m(t))$.

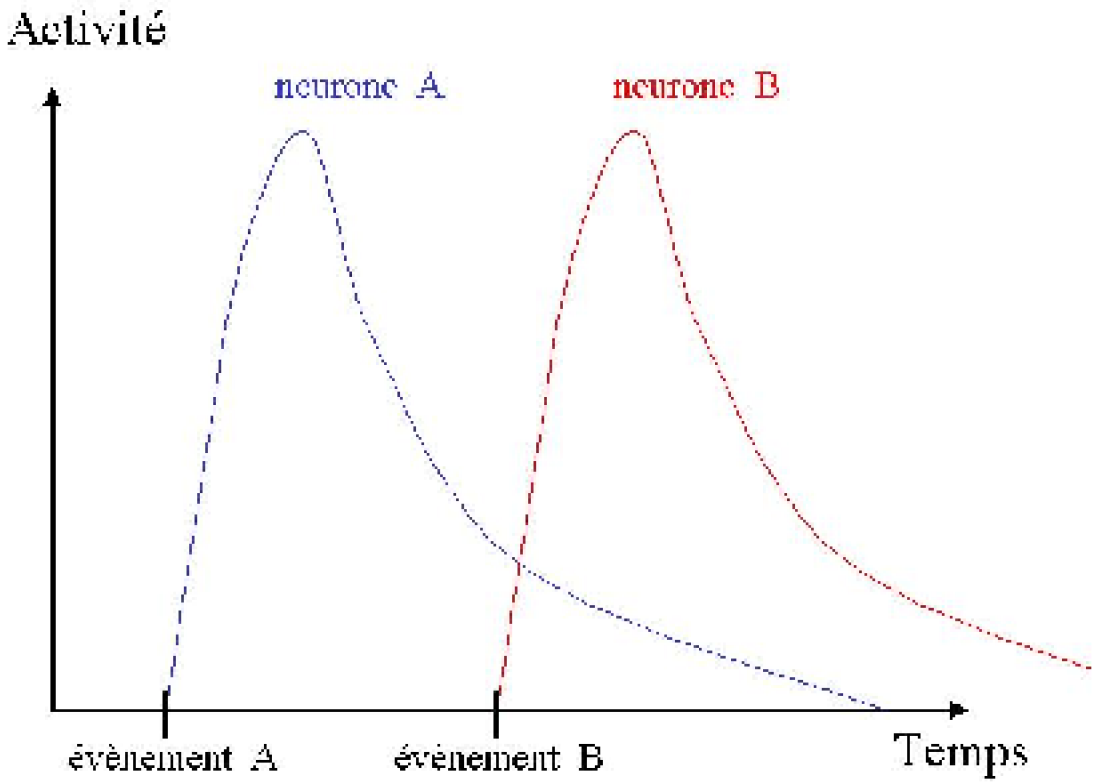


Figure 1.1 Coexistence des activités des neurones grâce au phénomène de trace.

La modélisation des neurones peut être réalisée, même à ce simple niveau, par différents modèles. Il n'est pas concevable d'avoir une approche qui serait automatiquement appropriée à tous les cas de figure. Un intégrateur à fuite se trouve particulièrement adapté aux problèmes nécessitant un grand nombre de neurones (Grethe et Arbib, 2001). Nous sommes bien dans ce cas, puisque la plupart des problèmes traités font appel à de nombreuses informations.

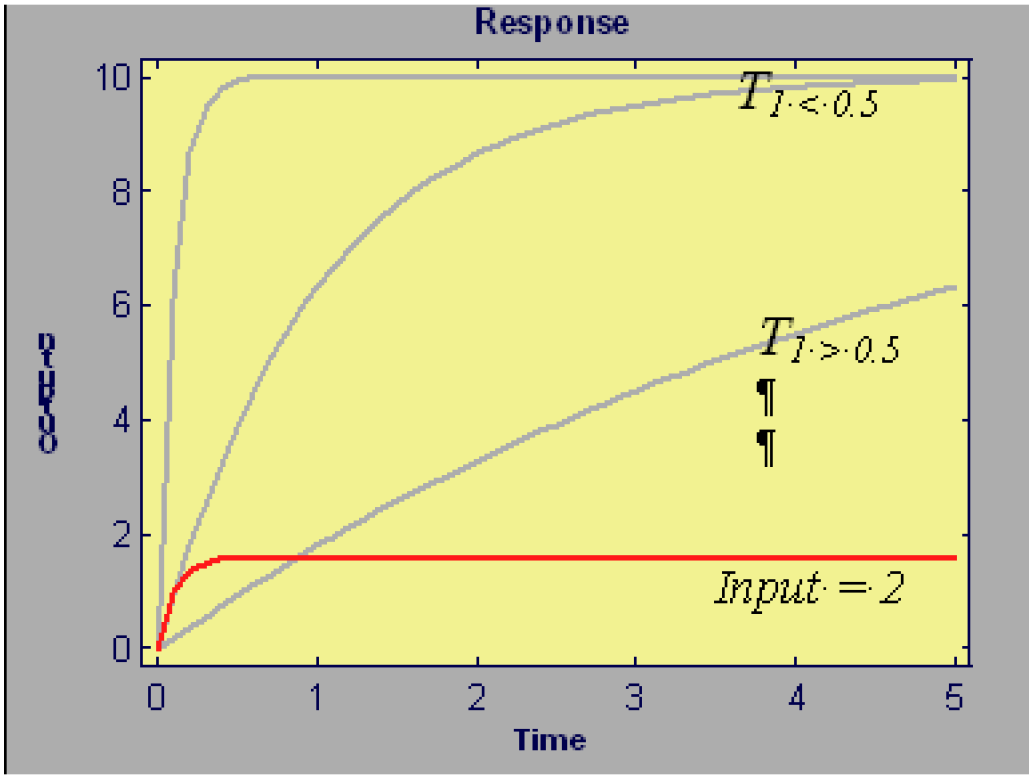


Figure 1.2 Réponse d'un intégrateur à fuite en fonction de trois constantes de temps différentes avec une valeur d'entrée de 10 (tracés grisés) et une valeur de 2.

III.3.1.2.Encodage du contexte

Dans ce modèle (Dominey et coll., 1995 ; Dominey, 1995 ; Dominey et Ramus, 2000), le contexte des événements passés est représenté grâce à des boucles récurrentes qui permettent à l'information présentée au moment *t* d'influencer la représentation des nouvelles informations au moment *t+1*. La couche State, qui correspond au cortex préfrontal avec ses connexions récurrentes locales et globales (Goldman-Rakic, 1987) est influencée par les entrées externes (Input) et les entrées récurrentes ($State_D$ [20]). Ainsi, l'utilisation des intégrateurs à fuite fournit une première mémoire des évènements sur une courte durée, alors que la boucle récurrent permet de retenir les informations pour des durées plus grandes. Cette architecture fournit une représentation du temps interne au modèle.

[20] State « D » fait référence à un encodage dynamique des informations contextuelles.

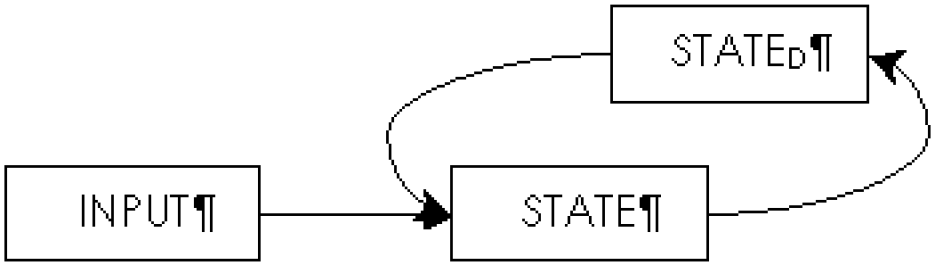


Figure 1.3 Architecture du réseau.

Dans (1.1) nous décrivons le « leaky integrator », *s*, qui correspond au potentiel de membrane ou activation interne de State. Dans l'équation (1.2) la sortie de State est générée par une sigmoïde, *f*, fonction de s(t). Le terme *t* désigne le temps, □*t* le « time step » de la simulation, □ est la constante de temps du « leaky integrator ».

$$s_i(t+\Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) s_i(t) + \frac{\Delta t}{\tau}\left(\sum_{j=1}^{n} w_{ij}^{IN} Input_j(t) + \sum_{j=1}^{n} w_{ij}^{RR} State_j(t)\right) \quad (1.1)$$

$$State_i(t) = f(s_i(t)) \quad (1.2)$$

$$sd_i(t+\Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) sd_i(t) + \frac{\Delta t}{\tau}(State_i(t)) \quad (2.1)$$

$$State_{Di} = f(sd_i(t)) \quad (2.2)$$

III.3.2.Apprentissage à partir d'un prototype moyen

Dans ce modèle, il n'existe pas d'apprentissage dans les connexions récurrentes. Initialement, un simple apprentissage associatif relie les différents états internes de « State » aux réponses appropriées. Le support neurophysiologique de cet apprentissage associatif correspond à la plasticité des synapses cortico-striatales sous influence dopaminergique (Dominey et coll., 1995). Avec cet algorithme d'apprentissage par renforcement, la structure temporelle peut être traitée indépendamment de la structure sérielle. Le modèle a déjà été testé avec succès dans des tâches de discrimination de langues à partir du rythme (Dominey et Ramus, 2000), ce qui suggère qu'il est effectivement sensible à la structure rythmique de la parole. Dans le cadre de cette thèse, cette modélisation sera étendue à un plus grand nombre de langues, ainsi qu'au traitement direct du signal acoustique, et à des problèmes d'identification portant sur des échelles temporelles différentes.

Pour pouvoir tester le potentiel du réseau pendant la phase d'identification, il faut étudier l'état des neurones des couches du réseau. Cette analyse est courante dans l'utilisation des réseaux récurrents (Elman, 1990 ; Cleeremans, 1993). Ceux-ci tirent avantage de la représentation développée dans ses unités qui sont copiées dans la couche de contexte $State_D$. Pour chaque point d'une séquence, ces motifs formés par les

unités de la première couche State codent la position de l'entrée courante dans la structure étudiée. La matrice des distances euclidiennes entre chaque vecteur d'activations permet de procéder à une analyse par regroupement (Cleeremans, 1993). Ce regroupement s'effectue en fonction du nœud de l'automate générant la grammaire apprise par le réseau.

Dans notre cas, les activations des unités sont stockées dans un vecteur de 50 unités, qui regroupe les activations enregistrées dans les deux couches State et $State_D$.

A chaque catégorie de séquences (langues, mots, attitudes prosodiques) correspond alors un vecteur prototype. Ce prototype est issu de la moyenne des activations des séquences du corpus d'apprentissage, pour un type de séquences données. Lors de l'identification d'une séquence, chaque distance avec le prototype d'une catégorie est calculée. La distance minimum indique alors la catégorie reconnue. L'évaluation d'un des états de la couche *State* se fait donc en deux temps :

1. Prototypage : Obtention d'un prototype de l'état induit par chaque catégorie ;
2. Evaluation de la distance entre un prototype et un état induit dans les couches State et StateD du réseau.

L'algorithme comporte trois boucles. Les deux premières passent en revue chaque matrice du corpus dont on veut connaître la catégorie qu'il décrit. Enfin, la troisième boucle permet de chiffrer la distance de la matrice étudiée avec chacun des prototypes des catégories. Finalement, le prototype présentant la plus petite distance indique la catégorie identifiée. Durant l'apprentissage, un prototype est créé à partir de la moyenne de l'ensemble des stimuli d'une même catégorie dans la base d'apprentissage.

```
Pour chaque catégorie I
    Pour chaque étiquette J du corpus
        Pour chaque catégorie J
            DISTANCE[ J ] = distance( MOTIF[ I, J ], MOYENNE[ I ] )
        Fin
        CATEGORIE_RECONNUE = position_minimum( DISTANCE )
        Si (CATEGORIE_RECONNUE = I
            Alors Incrémenter( CORRECT[ I ] )
    Fin
Fin
```

Algorithme 1.1 Evaluation de la première couche du réseau

La fonction **position_minimum**utilisée dans l'algorithme 1.1 renvoie la position du minimum.*MOTIF* désignele vecteur contenant l'activation de chaque neurone des couches State et $State_D$. *MOYENNE* désigne le tableau de la moyenne de la totalité des matrices de la base d'apprentissage pour une même catégorie. Le tableau *CORRECT* contient le nombre de séquences pour laquelle la catégorie a été correctement identifiée.

Chacun des problèmes étudiés pendant la partie expérimentale reprend ce principe. Cependant, des adaptions seront également examinées en fonction du problème étudié.

III.3.3.Différence avec les modèles récurrents « classiques »

L'utilisation d'un tel réseau récurrent temporel pour la représentation et l'apprentissage de séquences n'est pas nouvelle (Pearlmutter, 1995 ; Jordan, 1990 ; Elman, 1990). Ces réseaux intègrent la sortie précédente du réseau à l'entrée courante, pour déterminer leur état interne. L'apprentissage modifie les poids des connexions constituant le réseau de façon à ajuster au mieux la sortie obtenue à la sortie désirée. Cependant, avec les méthodes de rétropropagation récurrente couramment utilisées, la structure temporelle (et donc prosodique) ne peut être traitée indépendamment de la structure sérielle.

Afin de doter le réseau récurrent d'une sensibilité temporelle, chaque neurone est représenté par un intégrateur à fuite, dont un des paramètres est une constante de temps. Cette constante est directement reliée à la sensibilité temporelle du réseau. Nous verrons (Chapitre Six) qu'en modifiant cette constante, le réseau ne détecte plus les variations temporelles les plus rapides. En outre, l'échantillonnage des données ne dépend ni du type d'entrées (syllabes, mots, etc.) ni de la sortie désirée ou du processus d'apprentissage. Ainsi, les données en entrées peuvent avoir une durée arbitraire (Dominey et al., 1995 ; Dominey et Ramus, 2000).

Cet aspect du réseau est dénommé contrainte temporelle, elle implique que les informations sont fournies de façon séquentielle au réseau. Il n'est donc pas possible de traiter en parallèle des données qui ont eu lieu à des instant différents. De plus, la durée est représentée par le nombre de cycles pendant lequel un événement est présenté au réseau. Ainsi, si une consonne dure 60ms, elle est présentée 12 fois en entrée du réseau (soit 12 pas de 5ms).

La structure temporelle des entrées peut être traitée et représentée par le TRN avec une complexité spatiale et un coût informatique peu élevés. Effectivement, le réseau TRN fait appel a beaucoup moins d'unités que le modèle utilisé par Buonomano (2000) pour la discrimination de séquences simples. En outre, les algorithmes d'apprentissage inspirés de la rétropropagation du gradient (recurrent back-propagation et back-propagation through time) font appel à des ressources informatiques, qui ne sont pas biologiquement raisonnables (Pearlmutter, 1995).

Pour palier à ces défauts inhérents à une contrainte temporelle, les connexions récurrentes sont fixes, i.e. elles ne sont pas modifiées par l'apprentissage. Nous étudions une population de réseaux TRN, dont la seule différence est la répartition des poids des connexions. La sélection du réseau donnant les meilleures performances sur le corpus d'apprentissage permet de simuler un apprentissage modifiant le poids de ces connexions. Cette sélection ppourrait vraisemblement être améliorée par une sélection par algorithme génétique.

Dans le cas du réseau TRN, l'apprentissage s'effectue par renforcement lorsque la mémoire associative est utilisée (Ramus et Dominey, 2000) ou par un mécanisme d'apprentissage indépendant extérieur au modèle (Blanc et Dominey, 2003). Le réseau agit alors comme un mécanisme d'encodage des informations spatio-temporelles en informations spatiales.

Une autre différence avec les modèles classiques provient du type d'information traitée. Jusqu'à présent, le TRN traitait une information discrète. Dans ce contexte, un des neurones d'entrées du réseau représentait un événement, comme les consonnes ou les

voyelles (Dominey et Ramus, 2000). Dans cette thèse, nous posons l'hypothèse que ce réseau peut traiter une information continue dans le temps et dans l'espace, comme la fréquence fondamentale ou le spectre.

Les mécanismes de traitement des séquences temporelles, présentés précédemment, sont des systèmes d'analyse généraux, qui peuvent intervenir pour différentes modalités. Cependant, notre travail concerne uniquement la modalité auditive, c'est pourquoi nous proposons maintenant un rapide aperçu du traitement de l'acoustique par le système nerveux.

IV.Les séquences temporelles auditives : un bref regard sur le système auditif

L'organe sensoriel de l'audition comprend trois parties : l'oreille externe, l'oreille moyenne, et l'oreille interne. La première est extra crânienne, les deux autres sont comprises dans l'os temporal. L'oreille externe se compose du pavillon et du conduit auditif externe. La membrane appelée tympan sépare les parties externe et moyenne de l'oreille. La chaîne des osselets (marteau, enclume, étrier) relie le tympan à l'oreille interne, par la platine de l'étrier, qui est mobile dans la fenêtre ovale. Ces premières parties de l'oreille permettent une amplification du signal acoustique, avant son traitement, par l'oreille interne, qui contient les organes sensoriels. Notre intérêt se focalisera sur le fonctionnement de l'oreille interne, et sa capacité à représenter les signaux acoustiques.

IV.1.L'oreille interne

La représentation brute du signal acoustique dans le système nerveux auditif périphérique s'appelle *transduction sensorielle*. Ce processus comprend la transmission de vibrations à la cochlée, dans laquelle le signal excite différentes parties de la membrane basilaire en fonction des fréquences qui la composent. Le mouvement de la membrane en chaque point est converti en impulsions nerveuses transmises au cerveau par les fibres qui constituent le nerf auditif.

Il s'agit du premier stade de l'analyse sonore effectuée par le système auditif humain. Il se produit dans la cochlée ou organe de corti : le son y est décomposé en pattern neuraux distincts, représentant approximativement les différentes fréquences du signal. Le son est alors représenté par plusieurs milliers de cellules (hair cells) qui traduisent l'activité mécanique des vibrations en activité électrique. Chaque cellule ainsi que les nerfs auxquels elles sont reliées sont accordés sur une fréquence spécifique (Olshausen et O'Connor, 2002). A de faibles niveaux d'intensité, la bande de fréquence codée dans une fibre donnée est très étroite, alors qu'à de très fortes intensités la sélectivité fréquentielle d'une fibre particulière est plus faible (McAdams et Bigand, 1994).

IV.2.Une analyse spectrographique

Si l'on considère que le faisceau de fibres correspondant à des fréquences spécifiques représente un type de dimension spectrale et que le niveau moyen d'activité de différentes populations de fibres nerveuses représente un type de dimension intensive, on peut voir l'ensemble de l'activité, au fur et à mesure qu'elle évolue dans le temps, comme un spectrogramme neural du taux moyen. La configuration globale de l'évolution spectrale serait partiellement appréhendée par une telle représentation (McAdams et Bigand, 1994).

Actuellement, l'analyse spectrale et le spectrogramme sont les représentations les plus usitées pour le traitement des sons. Parallèlement, le système auditif est le plus souvent conceptualisé par une distribution spatiale, où une sous-population particulière des neurones auditifs est adaptée à un domaine donné de fréquences. Le profil des décharges des cartes tonotopiques produit une représentation neurale de l'analyse spectrale du stimulus. Cette représentation est basée sur des canaux : dans sa forme la plus simple, chaque neurone est caractéristique d'une fréquence. La vitesse des décharges du neurone indique alors la quantité d'énergie présente dans le stimulus pour une bande de fréquence (Cariani, Tramo et Delgutte, 1997). En utilisant, des méthodes statistiques, Lewicki et Arthur (1996) ont montré que les analyses auditives effectuées par la cochlée (en particulier la taille de la fenêtre d'analyse) dépendent des sons de l'environnement naturel.

Deux types de codages sont actuellement retenus :

1. Un codage spatial des fréquences par l'activité des neurones (codage par canal). Dans le cas du premier codage, un neurone ou un ensemble de neurones répond spécifiquement à une fréquence. Cette stratégie est principalement utilisée avec des cartes auto-organisatrices. Nous la reprendrons également avec le modèle TRN.
2. Un codage temporel dans le flot des décharges. Cette dernière représentation serait particulièrement adaptée aux nouveaux modèles issus des neurosciences en particulier les modèles avec des synapses dynamiques.

IV.3.Le paradoxe résolution-intégration

Un ton est la conséquence physique des fluctuations cycliques de la pression de l'air. Pour obtenir une estimation fiable de la fréquence, plusieurs cycles doivent être intégrés. Mais une longue période d'intégration signifie que la précision temporelle du mécanisme d'analyse (ou filtre) doit être réduite. En d'autre terme, il n'est pas possible de garder la plus grande précision à la fois dans le domaine temporel et fréquentiel. Le défi de tout système de traitement auditif est d'obtenir le meilleur compromis pour ces deux dimensions.

Ce paradoxe est appelé résolution-intégration [21] (Denham, 1999). Comment peut-on intégrer l'information donnée pendant une période assez longue, tout en n'excluant pas

les détails donnés par une résolution fine ?

La représentation des motifs acoustiques rapides par le cortex auditif est une question encore irrésolue. Effectivement, les performances des animaux et des êtres humains pour la discrimination de variations temporelles sont supérieures à ce que pourrait prédire les données neurophysiologiques (Lu, Liang et Wang, 2001). Les neurones corticaux intègrent les évènements acoustiques dans une fenêtre d'intégration temporelle (Wang, 2000). Dans le cortex auditif des marmousets, la taille de cette fenêtre varie entre 20 et 30 ms. Comment les neurones corticaux peuvent ils représentés des évènements qui ont une durée inférieure à la taille de la fenêtre (Lu et coll., 2001) ?

Les vocalisations peuvent être encodées par les motifs des décharges, formés par la distribution spatiale des populations de neurones. Ces stratégie de codage ont été retrouvées dans les nerfs auditifs (Sachs et Young, 1979 ; Young et Sachs, 1979) et le noyau cochléaire (Blackburn et Sachs, 1990). Beaucoup d'informations acoustiques sont codées dans les motifs d'activation des neurones auditifs. Cariani (1999) montre que la distribution des intervalles des décharges des fibres d'un nerf auditif, en réponse à un stimulus auditif, ressemble à la fonction d'autocorrélation du même simtulus. Une autocorrélation pourrait donc fournir un moyen au système nerveux pour effectuer une analyse de Fourier (Cariani et coll., 1997).

Les caractéristiques temporelles sont représentées par des décharges neuronales entre le cortex et la périphérie du système nerveux. Au contraire des nerfs auditifs, les neurones corticaux ne suivent pas exactement le décours des composants du stimulus. La capacité de suivi des neurones corticaux est limité à 20-30 ms. Pourtant, les composants les plus brefs ne sont pas perdus dans les représentations corticales. Certains neurones codent ces changements, en modifiant leur vitesse de décharge (Lu et Wang, 2000).

IV.4.Identification de séquences sonores

L'oreille permet d'obtenir les caractéristiques fréquentielles d'un son. A partir de cette représentation, le système nerveux doit pouvoir traiter l'ordre temporel de séquences composées d'un nombre limité de sons. Warren et Ackroff (1976) ont étudié l'identification de séquences sonores. Trois sons sont utilisés pour créer deux agencements (ABCABC… et ACBACBACB…). Le seuil d'identification de l'ordre est d'environ 450 à 670 ms par item. Ce seuil a été réduit 200 et 300 ms par item, lorsque les sons sont agencés à l'aide de cartes. L'agencement des cartes facilite donc le traitement séquentiel. Warren et Byrnes (1975) ont testé des séquences tonales de 4 items répétés, en jouant sur l'intervalle entre les différents sons (de 0.3 à 9 demi tons). L'augmentation de l'intervalle de hauteur ne réduit pas l'acuité du jugement.

Selon Watson et Kelley (1981) les sujets procèdent d'abord à une organisation globale des séquences de sons : « les caractéristiques d'ensemble du pattern » comme le contour de la hauteur et le rythme permettent de l'identifier et d'utiliser ensuite ces indices dans le traitement des détails d'un composant distinct. Lorsqu'elles forment des mélodies

[21] Ce paradoxe est aussi vrai pour la modalité visuelle (par exemple en photographie).

les séquence de tons ne sont pas perçues comme une série de hauteurs : elles possèdent une propriété émergente caractéristique du contour global.

Ces remarques suggèrent un traitement global, mais un traitement local est également nécessaire pour identifier chaque composant d'une structure sonore. Dans ce contexte, il paraît important de vérifier qu'un système connexionniste puisse traiter indifféremment des structures globales et locales, principalement si il est soumis à une contrainte temporelle.

V.Conclusion

Différents systèmes capables de traiter et d'analyser des séquences temporelles ont été présentés dans ce chapitre. La recherche en Informatique a fourni quelques modèles pour traiter des séquences discrètes. Cependant, ces modèles ne vérifient pas les propriétés connues du système nerveux. Dans ce contexte, des modèles sont apparus pour prouver qu'il était possible de traiter des séquences discrètes, avec des réseaux inspirés des neurosciences. Quelques uns de ces modèles sont en mesure d'apprendre ou de discriminer des séquences dont les structures temporelles sont distinctes. Cependant, les séquences testées sont souvent une vision simplifiée de la complexité des stimuli de notre environnement.

Nous proposons qu'au moins un de ces mécanismes, le réseau TRN, puisse être employé pour catégoriser différents types de séquences temporelles présentes dans notre environnement auditif, en respectant une contrainte temporelle. Ces séquences résultent d'une composante de la parole : la prosodie. Le chapitre suivant définit la prosodie avant d'aborder celle-ci à travers différents domaines qui forment un Continuum Temporel.

Chapitre Deux La Prosodie : Structure Temporelle de la Parole

« Il y a un vrai divorce entre la lange écrite et la langue parlée. Cela va donner naissance à un autre mode de communication » Raymond Queneau.
« Doukipudonktan, se demanda Gabriel excédé. Pas possible, ils se nettoient jamais. » Raymond Queneau, Zazie dans le métro, p. 1. « Que la musique soit un langage, par le moyen duquel sont élaborés des messages dont certains au moins sont compris de l'immense majorité, alors qu'une infime minorité seulement est capable de les émettre, et qu'entre tous les langages, celui-là seul réunisse les caractères contradictoires d'être à la fois intelligible et intraduisible, fait du créateur de musique un être pareil aux dieux et de la musique elle-même le suprême mystère des sciences de l'homme, celui contre lequel elles butent, et qui garde la clé de leur progrès » Claude Lévi-Strauss 1964, Le cru et le cuit Paris, Plon, p.26.

I.Première approche de la prosodie

I.1.Composantes perceptives de la prosodie

Les structures perceptives de la prosodie reposent sur 4 propriétés acoustiques :

1. Le rythme est la caractéristique principale de notre étude. Cette notion comprend le débit de parole, la longueur et la répartition des pauses, les allongements syllabiques, la durée de divers événement sonores (syllabes, phonèmes), etc.
2. L'intonation est souvent présentée comme le paramètre primordial de la prosodie. Elle contient le caractère chantant de la parole. L'intonation est principalement définie à partir de la fréquence fondamentale (notée F0). Elle est la conséquence de la vibration des cordes vocales et de la pression trans-glottique. Cependant, les indices physiques (F0) ne correspondent pas exactement au contour intonatif tel qu'il est perçu par l'oreille humaine.
3. Le volume sonore correspond au paramètre physique de l'intensité, c'est-à-direl'énergie contenue dans le signal au cours d'un intervalle de temps donné.
4. Le timbre est spécifique des instruments ou des voix. Il est perçu indépendamment de sa hauteur ou de son intensité. L'évolution du timbre est provoquée par la superposition des composantes harmoniques et non harmoniques durant l'émission du son. Une fois encore, cette dimension est difficile à expliciter physiquement puisqu'elle s'appuie sur l'ensemble des valeurs spectrales du signal.

Nous retiendrons des différentes composantes de la prosodie qu'elles ne se définissent pas uniquement à partir des caractéristiques physiques du signal. Néanmoins, la prosodie est un des constituants de la parole qui reste accessible à chacun d'entre nous sans connaissance particulière.

I.2.Rôles de la prosodie

La prosodie apporte une valeur ajoutée réelle dans nombre de contextes liés à la parole. Les récentes conférences intégralement dédiées à la prosodie [22] sont une illustration remarquable de ses applications futures et de sa présence « tacite » dans notre vie quotidienne.

I.2.1.Les attitudes et les émotions

Le rôle prépondérant de la prosodie concerne la transmission des différentes émotions à travers la parole. Les attitudes prosodiques permettent de connaître les intentions du locuteur et son état émotif, que ce soit pour des émotions brutes, peu ou pas contrôlées ou des émotions qui relèvent du sentiment et des attitudes.

La prosodie contribue aussi à influencer les auditeurs. Ainsi, la voix participe aussi à l'identification du locuteur sous l'angle sociolinguistique. Caradan (2001) décrit les traits prosodiques de l'enseignant, lorsqu'il doit réaliser ses objectifs pédagogiques, et donc se montrer en position dominante. Ainsi, le locuteur peut appuyer son discours par le biais de la prosodie pour mettre en emphase certains mots ou certaines phrases, de façon à

[22] Journées Prosodie 2001; International Conference on Speech prosody 2002 & 2004.

guider la compréhension de l'auditoire.

I.2.2.Fonction syntaxique

Le mode est une catégorie grammaticale traduisant le type de communication entre un locuteur et son auditoire. Le mouvement mélodique de la dernière partie de l'unité intonative signale la finalité (mouvement descendant) ou la continuité (mouvement ascendant). En Français, ce mouvement prosodique autorise la distinction entre les assertions et les questions [23] . Chez les adultes, la prosodie est toujours utile pour la compréhension des phrases. Dans certains cas, les ambiguïtés syntaxiques ne peuvent être levées qu'à partir de l'examen de la prosodie. Par exemple, la phrase « la belle ferme le voile » peut être interprétée de trois façons distinctes, suivant que le mot ferme est interprété comme un verbe ou mot [24] .

L'intonation joue un rôle de délimitation : elle aide à segmenter le discours en un nombre déterminé d'unités délimitant le frontières de certains constituants syntaxiques. Les pauses et la déclinaison de la fréquence fondamentale indiquent les frontières entre différentes phrases ou propositions. Par son pouvoir hiérarchisant, la prosodie contribue à distinguer les informations de premier et second plan. Cependant, la syntaxe ne permet pas de définir entièrement la prosodie, pas plus que la prosodie ne détermine totalement la syntaxe.

I.2.3.L'acquisition du langage

Les jeunes enfants ont recours à la prosodie pour comprendre le sens des phrases (Shady et Gerken, 1999), mais elle serait aussi une aide pertinente pour l'acquisition du langage (Morgan et Demuth, 1996).

Effectivement, une fois la cochlée formée autour de 4 mois ½ de gestation (Lecanuet, 1997), les fœtus réagissent aux ondes pures (Shahidullah et Hepper, 1992) et perçoivent les basses fréquences, qui constituent la prosodie. Les nourrissons préfèrent la voix de la mère, lorsqu'elle utilise le « parler bébé » dont les contrastes prosodiques sont accentués. Dès le plus jeune âge, les nourrissons sont sensibles aux variations prosodiques. Par exemple, les nouveau-nés ne peuvent pas utiliser de connaissances lexicales pour retrouver les mots au milieu du flot continu de parole. Il leur faudrait trouver d'autres indices pertinents pour segmenter le signal de parole en mots. Certains de ces indices pourraient être extraits de la prosodie (Johnson et Jusczyk, 2001).

I.2.4.Troubles de la prosodie

Certains enfants présentent des troubles liés au langage, qui pourrait être causés par un mauvais traitement de la prosodie. L'hypothèse a été émise pour deux catégories d'enfants avec des troubles du comportement.

[23] Par exemple, si une question est posée avec la structure syntaxique de l'assertion (« Et elle joint le geste à la parole ? » demanda Turandot. (Zazie dans le métro. R. Queneau).

[24] Une pause peut être inséré avant ou après le mot ferme pour éclaircir le sens de la phrase.

Les enfants SLI montrent des difficultés pour l'apprentissage de la parole : leurs premiers mots, leurs premières phrases sont produits plus tardivement que chez les autres enfants. Ils font aussi preuve de difficultés pour la maîtrise des structures syntaxiques. Les structures prosodiques les plus locales (i.e. à un niveau compris entre le syntagme et la syllabe) leur seraient difficilement accessibles, ce qui retarderait l'acquisition du langage. Le Chapitre 6 sera consacré à l'examen de cette hypothèse.

Les enfants autistes ont des difficultés pour élaborer une vie sociale. Ils présenteraient des troubles pour l'accession aux structures suprasegmentales de la prosodie. Ce dysfonctionnement leur empêcherait alors de percevoir convenablement les émotions et les intentions de leur locuteur, ce qui les bloquerait dans leurs relations sociales.

Ces observations supposent un double déficit du traitement de la prosodie : localement pour les enfants SLI et globalement pour les enfants autistes [25] . Quelles structures neurologiques pourraient être en cause, lors d'un tel dysfonctionnement ? Comment prosodie et neurologie peuvent-ils nous aider à éclairer le traitement de la parole par l'homme ?

I.2.5.Neurologie

Les investigations relatives au traitement neurophysiologique des indices prosodiques sont pour l'instant assez rares [26] . Dans une étude de potentiels évoqués, Steinhauer, Alter et Friederici (1999) ont montré que la présence d'indices prosodiques influence en temps réel le traitement d'un mot et la construction syntaxique d'un énoncé, à l'aide d'une expérience où les patrons intonatifs peuvent guider l'interprétation structurale initiale d'un énoncé. Cette étude illustre également les rapports étroits entretenus entre le traitement de la prosodie et celui de la syntaxe [27] . La compréhension des mécanismes cérébraux du traitement de la parole ne pourra s'effectuer qu'avec la mise en commun des connaissances issues de la linguistique, de la psychologie et de la neurologie, pour la résolution de questions spécifiques. M. Besson (2001) souligne la ***« nécessité évidente mais néanmoins difficile à réaliser, de collaborations interdisciplinaires »*** .

I.3.Enjeux de la prosodie pour l'ingénierie

Les paragraphes précédents ont exposé la prosodie sous des angles variés. Nombres de ces fonctions prosodiques ont été intégrées dans des solutions d'ingénierie pour le marché industriel de la parole.

[25] Ceci fait l'objet de rechercher en cours (The Role of Prosody in Language Processing for Children with Autism. Margaret Kjelgaard).

[26] A cause du développement récent des techniques d'imagerie d'une part, mais aussi des moyens de contrôle de la prosodie, d'autre part (Besson, 2001).

[27] D'autres chercheurs de l'institut Max Planck de Leipzig ont confirmé ces résultats (Jescheniak, Hahne et Friederici, 1998).

Dans ce contexte, les théories sur l'intonation sont au service du traitement de la langue. De nombreux logiciels de synthèse de la prosodie ont vu le jour, citons entre autres : Lacheret-Dujour et Morel, 2001; Morlec, Bailly et Aubergé, 2001 (*cf.* Morlec, 1997 pour une revue). La prosodie synthétique doit alors pouvoir être comparée à la prosodie naturelle en terme d'intelligibilité. Rilliard et Aubergé (2001) proposent un paradigme perceptif pour évaluer et diagnostiquer les performances fonctionnelles d'un générateur prosodique. L'intonologue est une aide précieuse pour l'informaticien lors de l'analyse et de la production des données concrètes. En retour, les informaticiens proposent des outils qui permettent l'analyse des données [28] .

Les modélisations informatiques de la prosodie ont besoin pour être opérationnelles d'un étiquetage prosodique [29] . La segmentation du continuum prosodique représente une opération extrêmement coûteuse en temps, et ce particulièrement pour la parole spontanée. Pour pouvoir faciliter ces opérations, des logiciels de traitement de la prosodie sont apparus (ProSig, PRAAT, etc.). Des outils et des méthodes automatisent les transcriptions prosodiques (Campione, 2001 à partir d'INTSINT défini par Hirst et Di Cristo, 1998).

I.4.Une autre description de la prosodie

L'orientation naturelle pour décrire la prosodie semble d'étudier successivement les différents paramètres de la prosodie. Seulement, ils ne sont pas toujours définis de manière isolée : l'intensité est souvent associée à la fréquence fondamentale, pour caractériser un mouvement prosodique, comme les accents. En outre, certaines notions comme le timbre sont encore difficiles à traiter [30] .

La prosodie sera alors abordée en fonction de ses domaines de définition : du plus global vers le plus local. Le rythme général d'une langue sera traité en premier. Ce paramètre se définit sur un ensemble de phrases, il est dit suprasegmental. Ensuite, les contours intonatifs seront examinés, avec un espace de définition intermédiaire (6 syllabes). Enfin, les dernières propriétés de la prosodie seront considérées pour les mots.

La perception des contours mélodiques s'effectue depuis une représentation globale vers une représentation locale que ce soit dans la langue ou dans la musique (Dodane, 2003). Le traitement global est en effet nécessaire avant un traitement de l'information plus analytique qui permet de décomposer une forme en ses éléments constitutifs [31] .

[28] Mertens, Auchlin, Goldman et Grobet (2001) décrivent un système de balisage pour traduire l'intention communicative. Zellner Kellner (2001) résume les enjeux associés à la simulation du rythme en synthèse de la parole.

[29] Ainsi que la majorité des applications du traitement de la parole.

[30] Projet de la qualité de la voix : VOQUAL'03.

[31] Dans la perception visuelle, l'information globale précède aussi l'information locale (Kimchi, 1992 pour une revue) et elle est identifiée plus rapidement que l'information locale (Navon, 1977 ; Proverbio, Minniti et Zani., 1998).

II.Le rythme en tant qu'indice suprasegmental

L'objet de ce chapitre est le rythme, en tant qu'organisation globale de séquences de sons. Selon Watson et Kelley (1981), les sujets procèdent d'abord à une analyse des caractéristiques d'ensemble du pattern comme le rythme. Chez l'être humain, le traitement de la parole souffre de la détérioration de la structure temporelle, mais supporte une diminution des informations spectrales (Shannon et coll., 1995).

Le rythme se définit à partir des impressions perceptives. L'idée principalement retenue pour décrire le rythme est de l'envisager comme une succession d'événements qui se distinguent par leur accentuation. Ces événements sont considérés alors d'après leur structure (la forme, l'agencement du tout et des parties), leur périodicité (le retour régulier de marqueurs identifiables comme les temps forts et faibles) et leur mouvement (leur succession séquentielle dans le temps).

Cette définition peut satisfaire à la fois le rythme tel qu'il est considéré en musique et en linguistique. Cependant, retrouver le rythme d'une pièce musicale, ou d'un discours ne semble pas avoir de points communs. Quelles sont les propriétés du rythme pour la musique et la parole ?

II.1.Pour la Musique

Le rythme peut s'entendre comme le sentiment du mouvement au cours du temps, perçu au travers des pulsations, du phrasé de l'harmonie et de la métrique (Lerdahl et Jackendoff, 1985; Large et Palmer, 2002). Les durées des événements rythmiques forment un motif temporel, au sein d'une séquence auditive.

La musique est marquée par des pulsations qui reviennent régulièrement et donnent le tempo. Celui-ci est initialement établi par des événements sonores, et il se maintient, même si d'autres événements entrent en contradiction avec ce tempo. Le plus souvent, un élément plus saillant, comme le martèlement de la grosse caisse ou le charleston permet de trouver le rythme. Tout retour périodique d'une différence acoustique va tendre à être interprété comme un temps fort de la structure métrique (Fraisse, 1974).

Quelques algorithmes automatiques permettent de déterminer le rythme ou la pulsation d'un morceau musical (Large et Kolen, 1994; Large et Palmer, 2002 ; Leman, Lesaffre et Tangue, 2001). Ils n'approchent cependant pas les performances de musiciens humains. Cependant, il n'existe pas un indice isolable susceptible de donner le tempo. Comment l'oreille humaine peut elle détecter et isoler ces indices au sein d'une pièce musicale comprenant de nombreux événements sonores ?

II.1.1.Les marques du rythme

Conformément à la théorie de l'analyse des scènes auditives, le groupement auditif précède généralement *l'extraction ou le calcul des propriétés ou attributs perceptifs.* Ainsi,

un brusque changement dans l'intensité d'un son permet de déduire l'adjonction d'un second son. Au contraire, si l'augmentation d'intensité se fait continûment, l'oreille humaine ne perçoit qu'une seule source sonore qui varie [32] .

Le rythme pourrait être marqué par les différences de certains paramètres acoustiques, comme la F0, ou l'amplitude. Le système auditif traite une transformation soudaine des propriétés acoustiques comme le début d'un nouveau signal. Quelques règles ont été décrites par A.S. Bregman (1994) pour expliquer la détection des événements sonores dans un flot continu, événements dont la succession donnerait naissance au rythme.

Régularité 1 : il est extrêmement rare que des sons n'ayant aucun rapport entre eux démarrent et s'arrêtent exactement au même moment.

Régularité 2 : Progression de la transformation [33] :

- Les propriétés d'un son isolé tendent à se modifier de façon continue et lentement.
- Les propriétés d'une séquence de sons issues de la même source tendent à se modifier lentement.

Régularité 3 : Lorsqu'un corps sonore vibre à une période répétée, ses vibrations donnent naissance à un pattern acoustique dont les fréquences des composants sont des multiples d'une même fréquence fondamentale. Effectivement, l'identité des sons est mieux perçue lorsque leurs harmoniques correspondent à deux fréquences fondamentales distinctes.

Régularité 4 : La plupart des modifications qui surviennent dans un signal acoustique affecteront tous les composants du son résultant de manière identique et simultanée.

Lorsque deux événements sonores sont perçus comme distincts, le système de traitement auditif crée l'illusion d'une pause sonore. L'expérience réalisée par Thorpe et Trehub (1989) en est une bonne illustration. Des motifs sont structurés de telle sorte que les trois premiers sons sont identiques. Les trois sons suivants possèdent soit une fréquence fondamentale, soit une structure spectrale (onde en scie ou sinusoïdale) soit une intensité différente (structure XXXOOO). Une pause peut-être insérée soit entre le troisième et quatrième son (XXX_OOO), soit entre le cinquième et sixième (XXXO_OO). Dans ce dernier cas, la pause insérée entre en conflit avec les autres indices qui permettent de découper le signal en deux flux distincts, ce qui entraîne dans ce cas un changement à l'intérieur d'un groupe. Les nourrissons détectent les modifications temporelles survenant dans un groupe mais pas celles entre les groupes, ce qui indique qu'ils regroupent le motif original de sons. Tout se passe donc comme si la perception du changement était équivalente à l'insertion d'une pause. Le motif désorganisant le motif

[32] L'œuvre musicale de Rimsky-Korsakov intitulée *le vol du bourdon* permet d'illustrer la nature de ce processus : les notes s'y suivent à une vitesse si rapide qu'elles forment un flux sonore unique et ne peuvent pas être entendues séparément.

[33] Par exemple la séquence de sons aigus (A) et graves (G) … A A G A G G A G A A G G A G … formera les 2 flux suivants : … A A _ A _ _ A _ A A _ _ A _ … et … _ _ G _ G G _ G _ _ G G _ G … Les tirets symbolisent la présence de silences dans chacun des flux.

standard (de 3-3 vers 4-2) est remarqué, contrairement au changement conservant cette structure (3-3 dans les deux cas). Les adultes perçoivent également la durée d'un silence situé entre deux groupes. Cette illusion de pause entre les groupes sonores est analogue à l'illusion de pauses entre les mots. Plus la discontinuité est grande, moins la détection des pauses entre les groupes est précise.

Chacune de ses règles fait appel à un certain nombre de paramètres acoustiques, dont les variations influent sur la perception du stimulus sonore. Ainsi, van Noorden (1975) examine la ségrégation des trilles. Les sujets doivent suivre le rythme de sons alternativement grave ou aigu (trille). Il étudie l'attention partagée entre les deux fréquences différentes et l'attention sélective au son grave uniquement. Les auditeurs ont des difficultés pour entendre un trille cohérent, lorsque la cadence et la séparation fréquentielle augmentent. En revanche, l'écoute des sons graves est possible pour de très petites différences fréquentielles. La différenciation entre deux événements sonores dépend au moins de la fréquence absolue de chacun des tons, et de l'intervalle fréquentiel entre les deux tons. Pour l'instant, il semble difficile de dire si ces règles et leurs seuils de distinction pour les paramètres associés sont innés ou issus d'une habituation à l'environnement.

II.1.2.Le traitement du rythme

La première étape consistait donc à identifier des éléments saillants dans le flux acoustique. La seconde doit pouvoir se servir de ces éléments pour retenir un motif rythmique, et saisir son retour périodique. Deux processus interviennent : le premier permet d'obtenir un codage hiérarchique des séquences rythmiques et privilégie les rapports simples entre les intervalles. Tandis que le second permet de reconnaître une même séquence jouée à des tempos différents.

II.1.2.1.Codage hiérarchique

Pour représenter une longue séquence d'événements, les sujets codent l'information contenue dans ces séquences par une hiérarchie d'opérateurs. Deutsch (1980) demande à des auditeurs musiciens de rappeler par écrit quatre types de séquences ou mélodies. Les performances obtenues démontrent que les auditeurs perçoivent et utilisent les hiérarchies pour mémoriser des séquences musicales. Effectivement, la mémorisation est facilitée lorsque la segmentation de la surface musicale coïncide avec la structure hiérarchique de la mélodie. Elle est beaucoup plus difficile dans le cas contraire.

Des auditeurs doivent suivre (explicitement ou implicitement) des séquences de bips de 150 ms séparés par des intervalles longs ou courts, ayant des rapports simples (2:1, 3:1, 4:1) ou complexes (3:2, 5:2, 4:3). Lors de la reproduction de ces motifs, les réactions chronométrées des auditeurs s'avèrent plus exactes pour des séquences ayant le rapport temporel le plus simple 2:1. En outre, les sujets savent reproduire des rythmes avec les autres rapports simples (3 :1, 4 :1) si ils sont imbriqués dans des structures d'ordre supérieur (Povel, 1981).

II.1.2.2.Un processus d'abstraction

Les adultes sont capables à partir d'une séquence, d'abstraire une structure rythmique qui restera identique malgré les variations de tempo. Vers l'âge de 2 à 3 mois, les nourrissons ont été testés avec des séquences [34] soumises à des variations de tempo (Demany, McKenzie et Vurpillot, 1977). Ils perçoivent certaines similitudes pour une même séquence exprimée avec un tempo différent, mais cette opération reste très difficile. Les étourneaux (Hulse, Page et Braaten, 1984) et les pigeons (Miller et Liberman, 1979) montrent également les mêmes performances. Cette abstraction est encore observée lorsque les événements qui ont créé le rythme se sont « évanouis », ou quand le rythme est soumis à de légères fluctuations.

La musique et son interprétation revêtent un paradoxe particulier. Alors que la structure temporelle de la musique et son organisation hiérarchique sont figées sur la partition musicale, il n'en va pas de même au niveau de l'interprétation musicale. ***« Le musicien virtuose joue avec le cadre rythmique qui lui est imposé, jusqu'aux limites de ce cadre, mais sans jamais le violer ; la musicalité est à ce prix et il semble bien que l'oreille n'apprécie pas les rythmes trop réguliers, qui lui paraissent non naturels »*** (Dodane, 2003, p. 34).

La régularité qui apparaît sur la partition n'existe donc dans la réalité de l'interprétation. Pourtant, elle constitue une réalité psychologique pour l'auditeur qui se montre capable d'extraire une pulsation. Le terme de « métrique » désigne ce processus d'abstraction. Même si l'interprétation musicale ne se montre pas parfaitement régulière, elle est donc perçue par l'auditeur comme métrique.

La section précédente a référencé les indices acoustiques, qui permettent de déduire le rythme, et donner quelques indications sur les propriétés des mécanismes, traitant la structure rythmique. Ces indices et ces propriétés se retrouvent-ils dans la parole ?

II.2.Pour la Parole

Obtenir le rythme de la parole semble encore plus ardu : si les musiciens s'entendent encore assez facilement sur la notion de rythme, aucune unité pertinente n'est mis en avant pour la parole. Les être humains sont capables de marquer le rythme de la parole et d'aligner d'autres propos au même rythme que ceux qu'ils entendent (Fraisse, 1974), même s'il est beaucoup plus difficile de battre la pulsation pour la parole que pour la musique (Drake, 2002). En outre, rien n'indique comment le système de contrôle, de production et de perception de la parole mesure le temps.

« Tout se passe comme si le sujet percevait successivement plusieurs groupes successifs d'éléments d'une manière semblable à celle dont nous lisons les lettres d'un texte, c'est-à-dire par des mouvements discontinus, avec de place en place, des arrêts pendant lesquels se produit la perception. » (Fraisse, 1967, p. 93).

[34] Les séquences de trois sons ont la structure 1-2 (X XX) ou 2-1 (XX X), celles de 4 sons ont la structure 2-2 (XX XX) ou 3-1 (XXX X).

II.2.1.Les marques du rythme

Plusieurs unités de la parole peuvent être candidates pour marquer la pulsation de la parole : les voyelles, la syllabe ou les accents. La durée de ces évènements peut aussi être considérée pour caractériser le rythme de la parole.

II.2.1.1.Les voyelles au cœur de la syllabe

Pour la parole comme pour la musique, le rythme peut être induit par la perception d'évènements réguliers. Le mouvement quasi-périodique de la mâchoire a donné très tôt un rôle prépondérant à la syllabe dans l'organisation rythmique de la parole (Fraisse, 1974). Mehler, Dupoux, Nazzi et Dehaene-Lambertz (1996) ont avancé l'idée que la perception des nourrissons est focalisée sur la voyelle [35] . Cette conception est soutenue par des remarques sur les propriétés acoustiques du signal de parole. Ainsi, la voyelle contient plus d'énergie, et possède une zone stable liée à son voisement. Les nouveau-nés se montrent plus attentifs aux voyelles qu'aux consonnes (Bertoncini et coll., 1988). Ils représenteraient la parole sous la forme d'une succession de voyelles, entrecoupés de bruits non analysés, les consonnes [36] .

Marcus (1975) propose un modèle de perception qui tient compte de cette périodicité. Les auditeurs sont capables d'ajuster l'intervalle de temps entre deux occurrences syllabiques (Voyelle *vs* Consonne-Voyelle) pour percevoir cette alternance comme isochrone. Il définit alors le centre de perception (ou Perceptual center) comme un point de repère psychoacoustique de la perception de la parole [37] (Morlec, 1997). Est-il possible de retrouver les voyelles (ou leur représentation partielle) à partir du signal de parole ?

Galvès, Garcia, Duarte et Galves (2003) suggèrent que les nourrissons utilisent des catégories grossières représentant les consonnes et les voyelles [38] . Ils introduisent donc une fonction de sonorité, qui reconnaît les motifs réguliers du signal acoustique à partir de l'entropie locale. Pellegrino (1998) définit également la voyelle à partir des basses fréquences. Le signal est d'abord prédécoupé en segments définis par des ruptures dans le signal [39] . Les voyelles sont alors identifiées à partir de l'énergie contenue dans les basses fréquences d'un spectrogramme.

En outre, la syllabe ne semble pas être le seul élément pouvant marquer un retour périodique dans la parole. Ainsi, le japonais voit son rythme lié à la mora (unité située entre la syllabe et le morphème ; Ladefoged, 1975).

[35] Modèle TIGRE : Time-Intensity Grid Representation.

[36] Cette hypothèse est à la base du travail effectué par F. Ramus dans sa thèse (1998).

[37] Ces repères se situent autour de l'établissement de la voyelle.

[38] Effectivement, le filtrage passe-bas simulant la prosodie, ne permet pas de retrouver très précisément les voyelles.

[39] Algorithme de divergence forward-backward (André-Obrecht, 1988).

II.2.1.2.Les accents

Le rythme peut aussi être considéré comme une succession d'unités faibles ou fortes (Cummins, Gers et Schmidhuber, 1999). Ces unités se traduisent sous la forme d'accents, qui ponctuent le discours. La définition linguistique exacte des accents dépend de la langue étudiée. Cependant, ils résultent d'une combinaison complexe de la hauteur, la durée, l'amplitude et des caractéristiques spectrales (Hirschberg, 1993). Ils se caractérisent également par un renforcement de l'énergie articulatoire, qui se traduit par une proéminence au niveau auditif. Ainsi, cette combinaison permet de mettre en relief la syllabe d'une unité (morphème, mot, syntagme) de la chaîne parlée. Ainsi, l'accent est mis en valeur, non seulement par ses qualités propres, mais aussi par les voyelles non accentuées de son voisinage qui perdent leurs traits distinctifs.

En Anglais, les mots marqués par un accent sont identifiables dans le contour de la fréquence fondamentale par des minima ou maxima locaux. Les mots perçus comme accentués sont souvent plus longs et ont une intensité plus élevé. Les mots qui ne sont pas accentués ne présentent pas ce genre de caractéristiques et leurs voyelles sont souvent réduites (Hirschberg, 1993).

La substance rythmique globale peut donc être influencée par l'intonation. Nous aborderons ce point plus précisément dans la section suivante, consacrée à l'intonation.

II.2.1.3.La durée

Les indices de durée classiques supposent généralement la donnée d'une segmentation, i.e. les frontières des unités dont la durée doit être mesurée. Les études conduites à ce sujet ont été motivés en majorité par la nécessité de modéliser la durée en synthèse de la parole.

Quatre classes regroupent les segments dont la durée a été caractérisée (Farinas, 2003) :

- Le **phonème**. Chaque phonème possède une durée intrinsèque, modifiée par un coefficient de rétrécissement, lui-même influencé par le contexte phonétique et l'environnement syntaxique.
- La **syllabe**. Elle se décompose en une attaque et une rime. La rime contient un noyau et une coda. L'attaque et la coda sont facultatives et constituées d'une ou plusieurs consonnes, le noyau ne comprend qu'une voyelle.
- Le **GIPC** (Groupe Inter Perceptual Center) est défini entre les points de perception (P-Center) de deux syllabes (Morlec, 1997).
- le **pied** se définit comme le plus petit groupement rythmique formé des syllabes inaccentuées précédant une syllabe accentuée pour les langues romanes. Le pied ïambique est défini par un motif : faible-fort, le rythme trochaïque par le motif inverse (fort-faible), dans ce dernier cas l'accent est porté par la première syllabe du mot.

Comment ces évènements sont-ils traités pour former une représentation du rythme de la parole ?

II.2.2.Traitement du rythme dans la parole

Les auditeurs adultes sont fortement sensibilisés à de nombreux aspects du rythme de la parole :

- Le rythme des accents (Lieberman, 1965). La perception des phrases anglaises est altérée lorsqu'un son réitéré avec un accent fort ou faible n'est pas aligné sur les marques linguistiques normales (Gleason et Bharucha, 1990). Les auditeurs perçoivent les accents qui tendent à marquer le début des mots (Cutler et Norris, 1988).
- Les motifs de durées entre les syllabes associées à différents mots (Smith, Cutler, Butterfield et Nimmo-smith, 1989).
- Les différences rythmiques globales entre certaines langues (Ramus, Nespor et Mehler, 1999).
- L'introduction de pauses même brèves (Peña, Bonatti, Nespor et Mehler, 2002). Celles-ci facilitent les opérations de segmentation, ce qui rend le champ libre pour que des sujets puissent extraire des règles d'abstraction sur l'organisation de pseudo-mots. Ainsi, les pauses permettent de retrouver l'organisation hiérarchique pour la musique (*cf.* II.1 ; Trehub et Taylor, 1998) et pour la parole également.

La perception du rythme dans la parole nécessite donc un certain nombre de traitement. Est-ce que ces mécanismes peuvent être utilisés par les nourrissons ?

Des bébés de 2 mois et demi sont capables d'organiser des sons successifs en fonction des intervalles de temps qui séparent ces sons (Demany et coll., 1977). Les intervalles de grande longueur étaient interprétés comme des pauses. Les nourrissons peuvent distinguer des langues issues de différentes classes rythmiques (Nazzi, Bertoncini et Mehler, 1998). La perception et la représentation de la parole par les bébés tendent à s'affiner. À deux mois, la discrimination entre l'Espagnol et le Catalan nécessite des capacités pour distinguer des propriétés prosodiques plus fines, notamment la position de la proéminence (Bosch et Sebastián-Gallès, 1997 cité dans Ramus, 1999). Enfin, à cinq mois, la discrimination Anglais/Américain suggère que l'enfant utilise des indices phonétiques et/ou prosodiques subtils (Nazzi, Jusczyk et Johnson, 2000). L'enfant acquiert de plus en plus d'informations sur sa langue maternelle, se focalise sur elle et s'intéresse moins aux autres langues. C'est ce que montre, à partir de l'âge de deux mois, l'absence de discrimination dès lors que la langue maternelle n'est pas concernée.

Le mouvement de l'intonation engendre lui aussi certaines sensations rythmiques, qui peuvent être étudiées dans leur globalité. Dans le cas des langues avec des tons lexicaux, la fréquence fondamentale n'est influencée que localement par la coarticulation des tons voisins. En revanche, pour les langues ne possédant pas de ton lexical, la tonalité est déterminée par la totalité de la phrase. Cette caractérisation globale peut-être obtenue par le calcul du coefficient de Hurst (Leavers et Burley, 2001).

Cette section a été consacrée aux caractéristiques globales de la prosodie exprimée donc sur l'ensemble des segments linguistiques. Par le fait, il s'agit principalement du

rythme, obtenu par divers indices comme les accents ou la syllabe, mais aussi par les variations de l'intonation. Cependant, l'intonation est le plus souvent décrite sur un ensemble plus réduit que le rythme. Certes, elle garde un aspect suprasegmental, mais ces propriétés se définissent à travers un domaine linguistique intermédiaire comme la phrase. Nous allons maintenant décrire un domaine plus local du Continuum Temporel de la prosodie.

III.L'intonation : Une approche suprasegmentale intermédiaire

III.1.Définition

L'intonation se comprend comme la mélodie de la parole, c'est-à-dire qu'elle est constituée d'une succession de sons, caractérisés par leurs rapports d'intervalles. Chaque son possède au niveau physique et acoustique une fréquence fondamentale (noté F0) qui correspond du point de vue de la perception à la notion de hauteur.

Il est important de retenir que l'intonation comme la mélodie pour la musique se caractérise par leur caractère séquentiel et leur dimension temporelle. Les sons qui composent la ligne mélodique entretiennent entre eux des rapports d'intervalles, c'est-à-dire une distance entre leurs hauteurs respectives. Cependant, cette définition tient plus de la mélodie que de l'intonation. Effectivement, la parole ne contient pas de sons isolés : en musique, le passage d'une note à une autre se fait de façon discontinue, alors que dans la langue, il se fait de manière progressive et continue, de telle sorte que les hauteurs ne peuvent être clairement délimitées et isolées par l'oreille (Dodane, 2003).

Nous distinguerons plusieurs points pour l'intonation. Nous commencerons par décrire les techniques employées pour obtenir l'intonation. Ensuite, le problème du traitement de cette information prosodique par l'être humain sera également abordé et nous terminerons par une description du « parler bébé », dont la particularité est d'exagérer les paramètres prosodiques de la parole, et plus particulièrement l'intonation.

III.2.Obtention de l'intonation

Après avoir brièvement évoqué les techniques permettant d'obtenir le corrélât physique de l'intonation, la fréquence fondamentale, nous aborderons le problème de l'obtention de l'intonation, qui résulte de l'interprétation des valeurs physiques par le système nerveux.

III.2.1.Valeurs brutes de la Fréquence Fondamentale (F0)

La fréquence fondamentale correspond à la fréquence de vibration des cordes vocales. L'estimation de la fréquence fondamentale continue d'être un champ de recherche actif

malgré le nombre d'algorithmes qui ont été proposés. Deux catégories distinguent les algorithmes fondés sur une représentation temporelle de ceux traitant le spectre du signal. Les méthodes temporelles utilisent la similarité du signal d'une période à l'autre pour identifier la période fondamentale. Le décalage observé entre le signal de départ et une répétition de ce signal indique la période du signal. Les méthodes spectrales s'appuient sur les harmoniques de la fréquence fondamentale. Hess (1983 ; 1992) et Hermes (1993) proposent des revues de ces algorithmes. Quelques exemples de méthodes sont [40] :

1. La fréquence instantanée (Abe et coll., 1995 ; Kawahara, Katayose, de Cheveigné et Patterson, 1999) ;
2. L'apprentissage statistique et les réseaux de neurones (Barnard, Cole, Vea et Alleva, 1991; Rodet et Doval, 1992; Doval, 1994) ;
3. Des modèles issus de l'audition (Duifhuis, Willems et Sluyter, 1982; de Cheveigné, 1991) ;
4. L'auto-corrélation (Boersma, 1993). La fonction d'autocorrélation est définie à partir d'un paramètre de décalage, et des valeurs du signal. Elle atteint son maximum, lorsque ce paramètre est nul. D'autres maxima sont présents dans le cas où le signal est périodique. Le décalage donnant le premier maximum correspond à la fréquence fondamentale, les suivants correspondent alors à ses multiples (harmoniques). Elle s'inscrit donc dans le cadre des méthodes temporelles.
5. Cross-corrélation ;
6. Sommation sub-harmonique.

III.2.2.Modèles de traitement de l'intonation

Avant d'aborder la perception de l'intonation chez les êtres humains, nous allons aborder le traitement de l'intonation. Ces modèles se subdivisent classiquement en deux catégories : phonétique et phonologique.

III.2.2.1.Modèles phonologiques de l'intonation

Les modèles phonologiques de l'intonation se fondent sur une représentation phonologique de la F0, c'est-à-dire un alphabet prosodique constitué de catégories discrètes avec une fonction linguistique, qui décrivent une unité particulière de l'intonation.

Une telle transcription de la prosodie et de l'intonation remonte jusqu'au siècle précédent. Les premières écritures se sont inspirées des notations musicales [41] .

De nos jours, le système de notation ToBI (Tones and Break Indices), dérivé des travaux de J. Pierrehumbert (1980), est sans doute le travail le plus influant dans ce domaine. Il repose sur un inventaire des tons : accents de hauteur (tons simples ou complexes), tons de frontières (alignés avec la fin d'une phrase), les accents de phrase

[40] La plupart de ces méthodes sont disponibles sur Internet.

[41] La thèse de C. Dodane (2003) explicite les différents types d'écriture.

(mouvement entre les tons de frontières et les accents de hauteur), plus un indice de rupture (ou pause). Les tons hauts sont notés **H** (« High ») et **L** (« Low ») marque les tons bas (Silverman et coll., 1992). Les tons situés à des frontières de phrases sont indiqués par le symbole **%**. Dans ce modèle, l'intonation est seulement déterminée par des composants locaux. Elle est décrite au niveau de la phrase intonative, constituée de phrases intermédiaires. En outre, ToBI inclut une notation des ruptures entre les mots, décrivant ainsi les séparations entres différentes phrases intonatives, et différentes phrases intermédiaires (Sun, 2002).

Cette notation avait été initialement développée pour l'Anglais, et les langues germaniques, puis elle a subit quelques modifications pour être adaptée à des langues aussi variées que le Japonais ou le Français, même si cette utilisation est controversée (Martin, 2001).

D'autres travaux ont appuyé leur modèle sur la perception de l'intonation. Seuls les mouvement de l'intonation qui sont perçus sont conservés. Une procédure itérative permet d'éliminer les contours de F0, qui n'ont pas d'influence sur la perception de la hauteur. Ensuite, une procédure de standardisation dresse l'inventaire des mouvements intonatifs. Enfin, une grammaire intonative est construite pour exprimer les combinaisons possibles de ces mouvements. Contrairement à l'approche précédente, l'unité intonative minimale est le mouvement de F0 au lieu des tons (Sun, 2002).

Ces modèles fournissent une description de l'intonation à un niveau cognitif. Cependant, ils requièrent l'expertise d'un linguiste connaissant la langue à étudier et ne proposent pas de modèle universel satisfaisant (Buhmann et coll., 2000).

III.2.2.2.Modèles phonétiques de l'intonation

Les modèles phonétiques de l'intonation considèrent les valeurs acoustiques de la fréquence fondamentale. Contrairement aux représentation précédentes, ils ont le plus souvent recours à une représentation continue pour décrire l'intonation (Buhmann et coll., 2000)

Dans ce contexte, les approches non paramétriques apprennent directement les valeurs de F0, à partir d'un ensemble de valeurs pour chaque syllabe. Un ensemble de paramètres prosodiques est extrait de la cinématique des trajectoires de F0 (Morlec, 1997). Ainsi, elles tiennent compte d'un nombre plus restreint d'informations linguistiques. Buhmann et coll. (2000) et Traber (1992) ont utilisé un réseau de neurones récurrent de type Elman (1990) avec un algorithme BPPT (Backpropagation through time) pour générer l'intonation. Traber (1992) incorpore une double mémoire à son réseau (métaphore spatiale en entrée et rétroaction (Morlec, 1997). Pour Buhmann et coll. (2000) 25 paramètres (indices de rupture entre les mots, proéminence des mots, nombre de mots dans une phrase, etc.…) permettent de prédire cinq valeurs équidistantes de F0 par syllabes, pour six langues. Ils montrent ainsi que ces derniers modèles sont bien indépendants de la langue (Sun, 2002).

A l'opposé, les modèles paramétriques requièrent une transformation de valeurs de F0. Le modèle Tilt (Taylor, 2000) emploie une représentation continue de la fréquence fondamentale. Dans ce cadre, les tons décrits par les modèles phonologiques

apparaissent comme des points particuliers de l'espace utilisé. Cette représentation est à la fois phonologique et acoustique. Les connaissances linguistiques proviennent des étiqueteurs du corpus, qui déterminent la présence d'un événement intonatif comme un accent. Ensuite, cet accent est représenté par un son type (ascendant, descendant ou une combinaison des deux) ainsi que par su durée et la somme de l'amplitude et de la montée et de la chute de F0. Le paramètre du type de l'accent est calculé automatiquement à partir des amplitudes et des durées de la montée et de la chute de F0. Dans ce cas, les étiqueteurs doivent indiquer la position et la durée d'un accent dans le signal. Ensuite, un système automatique retrouve les évènements intonatifs à partir de la représentation acoustique du signal.

Le système INTSINT [42] (Hirst et Di Cristo, 1998) décrit l'évolution de la F0 sous la transcription orthographique ou phonétique. Les mouvements mélodiques peuvent être figurés par des flèches, telles que ↑ pour indiquer une montée et ↓ pour une descente. Les flèches désignent les pics ou les vallées de la courbe intonative, qui sont jugés comme les moments les plus informatifs du point de vue de la perception et de la production. Il s'agit d'une transcription formelle et inversible de la structure mélodique.

L'algorithme de modélisation automatique (MOMEL) permet d'extraire automatiquement une séquence de points cibles constitués par des couples de valeurs <F0, temps>. Chaque point cible est codé par un symbole indiquant soit un ton absolu (Top, Bottom ou Mid), soit un ton relatif (Up, Down, Same, Higher ou Lower).

Les données brutes de F0 sont divisées en deux constituants : un facteur micro-prosodique correspondant aux variations à court terme de F0, qui sont conditionnées par la nature des phonèmes, et un facteur macro-prosodique, correspondant aux variations à long terme de l'intonation (indépendamment des phonèmes utilisés). Une courbe continue lisse obtenue à partir de fonctions « splines » quadratiques relie les points cibles et caractérise cette composante macro-prosodique. Elle correspond à des variations perceptives de F0, indiquant ainsi le profil suprasegmental qui définit globalement l'intonation.

Un outil d'analyse/resynthèse (technique PSOLA, Hamon et coll. 1989), utilisant les points cibles ainsi détectés, permet la synthèse de la courbe originale à partir de la courbe modélisée. Ce procédé est utilisé pour la validation perceptive de la stylisation automatique de la F0. Cet algorithme a été évalué pour différentes langues européennes à partir du corpus MULTEXT de EUROM1. Pour les enregistrements Anglais et Français, (soit une heure et demi), seulement 5% des points cibles ont du être corrigés pour satisfaire l'oreille humaine (Campione et Véronis, 1998).

Ce modèle peut être considéré comme un système combinant les aspects phonologiques et phonétiques. Une description de bas niveau du signal est d'abord employée pour obtenir une description phonologique de l'intonation. Quatre niveaux permettent donc de décrire la prosodie : tout d'abord, 1) le niveau acoustique procure les valeurs de F0, ensuite 2) le niveau phonétique est constitué des points cibles repérés par MOMEL, puis 3) le niveau phonologique de surface symbolise une séquence de

[42] « INternational Transcription System for INTonation ».

segments tonaux, enfin 4) le niveau phonologique sous-jacent spécifie les relations entre les tons. Seul le dernier niveau ne peut être obtenu automatiquement. Les trois premiers niveaux sont de plus indépendants de la langue.

Les modèles séquentiels de description de l'intonation se distinguent des modèles superpositionnels, par leur traitement de l'information. Les modèles superpositionnels, comme le modèle de Fujisaki, décrivent l'intonation suivant différents domaines temporels (phonèmes, syllabes, mots, phrases). Ils cumulent ensuite ces différents niveaux pour obtenir une représentation complète de l'intonation. Les modèles séquentiels (Taylor, 2000 ; Traber, 1992 ; Buhmann et coll., 2000) décrivent les valeurs de F0 sous forme de séquences de valeurs ou de mouvements de F0, en respectant l'écoulement du temps (Buhmann et coll., 2000).

III.2.2.3.Génération de l'intonation

Deux types de modèles de production de F0 sont communément employés (Sun, 2002). L'un s'appuie sur un ensemble de règles façonnées par des experts linguistiques. Cette première catégorie tend à produire des intonations peu variées. L'autre utilise des algorithmes d'apprentissage qui établissent des règles ou des relations à partir de données enregistrées et étiquetées par des labels prosodiques. Cette dernière catégorie dépend fortement des données disponibles pour la tâche souhaitée, mais aussi du type d'apprentissage employé (arbres de décisions, réseaux de neurones).

III.2.3.Perception de l'intonation

Le mécanisme permettant d'obtenir la fréquence de tons complexes est étudié depuis 150 ans. Les modèles d'autocorrélation temporelle sont apparus dans les années 1950 (Cariani et coll., 1997). La découverte dans les années 60 des régions de dominance de fréquence pour la hauteur dans les zones cérébrales a entraîné un regain d'intérêt pour les modèles fondés sur des motifs spectraux. Aujourd'hui, la représentation de la hauteur au niveau du nerf auditif sous la forme d'une population d'intervalle ressemble aux fonctions d'autocorrélation pour le même stimulus. L'autocorrélation permet de distinguer des voyelles à partir de cette distribution sur de courts intervalles de 5 ms. Ainsi, une description entièrement temporelle suffit à caractériser le timbre de sons stationnaires (Cariani et coll., 1997).

Les mécanismes permettant d'établir l'intonation sont encore aujourd'hui source de débat, comme pour le phénomène de restitution de la fondamentale absente [43] , bien connu en psychoacoustique et qui résulte en toute probabilité d'un traitement central et non périphérique [44] (Houtsma et Goldstein, 1972). Toutefois, quelques techniques émergent pour pouvoir représenter la fréquence fondamentale. Pour notre approche, nous privilégierons les représentations sous forme de spectrogramme, qui sont les

[43] La fréquence fondamentale n'apparaît pas dans le valeurs physiques, mais est déduite des valeurs des harmoniques qui sont accessibles dans le signal.

[44] Il se produit au niveau des gyri de Heschl de l'hémisphère droit (McAdams & Bigand, 1994).

techniques les mieux maîtrisées jusqu'à présent, et se trouvent être également adaptées au réseau récurrent TRN. Il nous reste maintenant à approfondir les connaissances acquises sur les mécanismes qui permettent d'extraire des informations de l'évolution de l'intonation.

III.3.Traitement de la hauteur

Afin de mieux cerner la façon dont s'effectue le traitement de la F0, nous présenterons des études comportementales chez l'adulte et des expériences conduites pour isoler les structures cérébrales responsables du traitement de la hauteur, avant d'aborder quelques expériences menées avec des nourrissons.

III.3.1.Traitement de la hauteur chez les adultes

Des auditeurs doivent juger si deux sons ont une hauteur identique ou différente (Deutsch, 1970 ; 1975). Les deux sons durent 200 ms et sont séparés par une pause de 5 secondes. Quatre conditions sont définies (les pourcentages indiquent le nombre de réponses correctes) :

1. Absence d'interférence, soit un silence de 5 secondes : 100 % ;
2. L'intervalle comporte six sons différents : 67.6 % ;
3. Une liste orale de six chiffres est énoncée pendant la pause : 97.7 % ;
4. Les sujets doivent énumérer les six chiffresLes chiffres sont rappelés avec une performance identique lorsque les sons sont absents. : 94.4 %.

Les interférences de sons sont plus dommageables lorsqu'elles proviennent du même registre de hauteur que les deux sons à comparer. Tout semble indiquer que les sujets se fient à une représentation du contour de la mélodie qu'ils viennent d'entendre et n'accordent que peut d'importance à la hauteur exacte des intervalles. Pour les adultes, une mélodie transposée ressemble à la mélodie originale. Au contraire, une mélodie avec le contour entravé se distingue de la mélodie d'origine. Les changements préservant le contour sont à mi-chemin (Trehub et Trainor, 1994). Les hauteurs absolues des notes sont utilisées pour la représentation de ces séquences chez les oiseaux chanteurs (Hulse et coll., 1990) et chez les singes (D'Amato, 1988).

Quelles structures cérébrales peuvent être à l'origine de ce traitement ?

III.3.2.Aspect neurologique du traitement de la hauteur

Les mécanismes neuraux permettant le traitement de la hauteur sont mieux connus que ceux traitant le rythme. La perception de la hauteur serait située dans l'hémisphère droit. Le cortex primaire effectue la première analyse acoustique de tous les signaux, et les aires associées sont responsables du traitement de signal d'ordre supérieur. Le cortex préfrontal droit ferait parti d'un réseau distribué pour maintenir en mémoire la hauteur. Cette hypothèse est soutenue par des données anatomiques, et par des études sur des lésions (gyrus temporal supérieur droit, et lobe frontal droit) qui créent des troubles pour la

rétention de la hauteur, ainsi que par une étude en TEP de Zatorre, Evans, Meyer et Gjedde (1992).

Des lésions du cortex auditif primaire droit (mais pas le gauche) troublent la perception de la fréquence fondamentale manquante (Zatorre, 1988). Wetzel, Ohl, Wagner et Scheich (1998) ont trouvé que l'hémisphère droit des gerbilles était spécialisé pour le traitement des modulations de fréquence.

Johnsrude, Penhune et Zatorre (2000) ont montré que des lésions des aires centrales dans l'hémisphère droit entraînaient un déficit spécifique pour juger de la direction d'un changement de hauteur. Les patients pouvaient compléter la tâche, mais avec beaucoup plus de difficultés (un seuil quatre fois supérieur). Dans ce cas, les patients utiliseraient l'hémisphère gauche, encore intact, mais avec une sensibilité pour les changements de hauteur beaucoup moins importante.

Jusque-là, nous avons présenté comment les adultes perçoivent la hauteur, qu'en est-il pour les nourrissons ?

III.3.3.Traitement de la hauteur chez les nourrissons

Dès leurs premiers jours, les nourrissons sont sensibles aux motifs mélodiques. Vers 2-3 mois ils sont sensibles à différents contours intonatifs associés à la même syllabe « pa ». Dès 4 mois et demi, les bébés préfèrent des extraits tirés des menuets de Mozart segmentés correctement (Jusczyk, 1989; Jusczyk et Krumhansl, 1993), car ils écoutent plus longtemps les passages interrompus par des pauses situées aux limites des phrases plutôt qu'à l'intérieur. Une phrase non résolue du point de vue tonal ne peut pas être considérée comme achevée.

Certains chercheurs postulent que, chez les nourrissons, le traitement local l'emporte sur le traitement global (Aslin et Smith, 1988), alors que d'autres penchent pour l'approche contraire (Morrongiello, 1988).

Pour le savoir, Trehub et Trainor (1994) ont étudié la capacité de bébés de 6 à 7 mois et d'adultes à détecter des modifications dans une mélodie composée de 6 sons. Les trois variantes débutent et se terminent sur les mêmes notes, mais la première est une transposition de la mélodie standard dans une autre tonalité, la seconde comprend des changements de notes qui préservent le contour et la dernière, des changements qui ne préservent pas le contour. Les deux premiers types de changement apparaissent aux bébés comme des reprises supplémentaires de la mélodie d'origine. Leur résultat est sensiblement conforme à la stratégie de traitement global, où l'information concernant le contour de hauteurs domine celles de la hauteur absolue et des intervalles. Ils traitent les contours de manière globale, en ignorant les détails de ces contours.

Les notes distantes d'une octave (rapport 2:1) sont en quelque sorte perçues comme équivalentes et cela, même par les nourrissons (Demany et Armand, 1984). L'échelonnement des hauteurs musicales, perçues sur la base de fréquences linéaires, est une autre caractéristique universelle. Ces caractéristiques reflètent probablement des contraintes de traitement du système perceptif [45] .

[45] Des modèles fondés sur l'autocorrélation permettent d'expliquer le caractère particulier de l'octave (Cariani, 1999).

Le traitement de contour de hauteurs fonctionne peut-être chez les nourrissons comme un important mécanisme d'organisation perceptif qui dirige leur segmentation des motifs auditifs complexes. D'autres mécanismes (groupement temporel, motif rythmique) en font également partie, mais avec moins de poids.

Les processus de groupement perceptif interviennent en fait déjà chez les nourrissons. Le traitement de relations des hauteurs et le traitement temporel ont les mêmes caractéristiques chez l'enfant et chez l'adulte. Le traitement de relation de hauteur semble être une prédisposition liée à la nature même du cerveau humain. Les nourrissons traitent l'information auditive globalement, en extrayant les contours de hauteurs des mélodies ou des expressions verbales, tout en ignorant de nombreux détails à l'intérieur des contours (Trehub et Trainor, 1994).

Nous venons d'examiner une partie des processus du traitement de la hauteur chez le nourrisson, ce mécanisme est impliqué lors des premiers contacts de l'enfant avec sa langue maternelle, par le biais du « parler bébé », que le paragraphe suivant va décrire plus longuement.

III.4.Le « parler bébé », langage adressé à l'enfant

L'intonation intervient dans une langue universellement répandue, destinée aux enfants dénommée « parler bébé » ou Langage Adressé à l'Enfant, (LAE, ou CDS pour Child Directed Speech). Deux hypothèses sont retenues pour expliquer l'émergence du langage chez l'enfant. Soit le langage est inné (hypothèse de la grammaire universelle), soit les entrées acoustiques de la parole induisent la structure de la parole. Dans ce dernier cas, le « parler bébé » permet de focaliser l'attention des enfants.

La parole destinée à l'enfant comporte une hauteur plus élevée, une gamme plus riche de hauteurs, des transpositions de hauteur plus souples, des contours de hauteurs plus simples, un tempo plus lent, des rythmes plus réguliers, des expressions plus courtes et des répétitions plus nombreuses.

Le « parler bébé » possède une grande musicalité définie par cinq ou six contours prototypiques qui ne se retrouvent pas dans la parole adressée à l'adulte [46] . Ces contours sont composés d'intervalles ayant des rapports fréquentiels simples, facilement décodables par le système perceptif (Burns et Ward, 1982 ; Rakowski, 1990) : les tierces, les quartes, les quintes et les octaves (Fernald, 1976) y sont prédominantes. Contrairement à la parole chez l'adulte, le langage adressé à l'enfant n'utilise pas de saut d'octave abrupte, mais garde des contours continus de l'intonation, alors même que les variations de l'intonation sont importantes.

La stimulation verbale maternelle semble s'accorder intuitivement aux capacités de traitement du nourrisson en général [47] . Les nouveau-nés préfèrent la parole prononcée avec des grandes variations de F0 et des contours ascendants rapides (Fernald, 1985 ; Fernald et Kuhl, 1987 ; Fernald, 1989). Ils préfèrent la voix de la mère en « parler bébé »

[46] 77% de ces contours sont caractéristiques du « parler bébé ». Les contours montants servent à attirer l'attention du bébé, tandis que les contours en cloche, dont la tessiture est plus large, servent à maintenir son attention.

dont les caractéristiques prosodiques sont fortement exagérées [48] . Les pics de F0 sont significativement plus élevés dans le parler bébé (Fernald et Kuhl, 1987).

Cette préférence n'est maintenue que pour la F0, lorsque les différents paramètres prosodiques sont isolés. Ce paramètre reste donc primordial par rapport au rythme, aux formants et à l'intensité pour le nouveau-né (Fernald et Kuhl, 1981).

Les enfants font très attention aux syllabes fortement accentuées, et ignorent celles qui ne le sont pas (Gerken, 1994). Lors de la production des mots isolés, les enfants Anglais omettent les syllabes faibles des rythmes ïambiques, mais moins souvent pour les rythmes trochaïques. Les enfants sont sensibles à la structure prosodique de la parole, à la fois au niveau des mots, mais aussi des phrases.

A l'âge de 9 mois, les nourrissons maîtrisent toutes les configurations intonatives, telles que l'interrogation, l'assertion ou l'exclamation. Leurs productions sont correctement interprétées par les parents et même par des auditeurs extérieurs au cercle familial qui leur attribuent une fonction (phatique) ou une modalité précise (question, énonciation). Et c'est uniquement l'intonation qui permet cette différenciation (Konopczynski et Tessier, 1994 cité dans Dodane, 2003). A onze mois se forment les moules prosodiques. Ces moules mélodiques formeront le squelette pour des structures syntaxiques plus complexes. Entre 18 et 20 mois, deux mots sont combinés entre eux, en respectant les contours prosodiques.

Cette richesse prosodique permettrait aux enfants de se sensibiliser à la structure du langage. Ils useraient donc des caractéristiques globales fournies par le rythme, et également par l'intonation, avant d'aborder des détails plus fins du signal de parole, dont la portée est uniquement locale, c'est-à-dire réduite à une syllabe. Nous avons examiné la contribution de l'intonation au sein de la prosodie. Celle-ci est distribuée sur l'ensemble de la phrase. Elle garde donc un caractère suprasegmental comme le rythme, mais elle se définit le plus souvent pour des durées plus courtes.

Nous allons maintenant catégoriser les propriétés de la prosodie, qui sont situées à l'autre boût du Continuum Temporel. Elles sont locales, et se définissent sur un segment linguistique court, comme la voyelle, la syllabe, voire le mot.

IV.La prosodie locale

Les bébés sont capables de distinguer deux langues étrangères, lorsque celles-ci sont suffisamment distinctes, ce qui montre qu'ils apprennent certaines propriétés de leur langue maternelle. Cependant, la nature de la représentation prosodique établie par les bébés est encore très mal connue. En particulier, les propriétés prosodiques peuvent être

[47] Kuhl et coll. (1997) indiquent que les voyelles et les consonnes sont prononcées par la mère de manière plus distincte lorsque l'enfant commence à parler (hyperarticulation), que lorsqu'il est plus jeune ou plus âgé (Ratner, 1984).

[48] Et ce même pour les prématurés (Nowik-Stern et coll., 1998).

définies sur un champ linguistique local (i.e. segmental). Ces composantes sont l'intonation sous forme d'accent ou encore de pics, les pauses, qui constituent des composantes purement prosodiques et le timbre.

Nous traiterons d'abord de ce que marquent ces indices purement prosodiques, soit des frontières, soit des particularités syntaxiques. Puis, nous aborderons les indices spectraux, situés entre l'acoustique et la prosodie, avant de décrire la perception de ces indices brefs.

IV.1.Détermination des frontières

Les indices prosodiques locaux peuvent marquer au moins deux types de constituants linguistiques enchâssés : les phrases et les mots. Le traitement automatique de la prosodie s'est également intéressé à la prédiction de ces indices (en particulier les indices de ruptures) dans le discours.

IV.1.1.Les frontières des phrases

A la différence du texte écrit, le langage oral ne contient pas de segmentation explicite fiable, comme des pauses entre les mots. Les bébés perçoivent un discours continu. Le découpage en unités syntaxiques s'effectue par un effet de compétition et de segmentation explicite. Plusieurs source d'informations sont disponibles : les marqueurs prosodiques, les contraintes phonotactiques, et les régularités statistiques.

A 4 mois et demi, les nourrissons préfèrent les extraits de parole, qui sont interrompus à la limite des phrases (Jusczyk, 1989). A 9 mois, ils montrent la même préférence avec des extraits, soumis à un filtre passe-bas (Jusczyk et coll., 1992). Les frontières prosodiques sont marquées par divers indices pour distinguer différentes propositions ou phrases :

- Les **segments phonétiques** (voyelles) tendent à être **allongés** aux frontières de phrases ou de propositions (Morgan et Demuth, 1996 : Klatt, 1975, 1976 ; Klatt et Cooper, 1975 ; Luce et Charles-Luce, 1983 ; Bernstein-Ratner, 1986).
- L'**intonation est descendante** à la fin des structures syntaxiques majeures, en particulier pour les phrases.
- **Les pauses** apparaissent aux frontières syntaxiques majeures, plutôt qu'aléatoirement dans la phrase, la longueur des phrases reflète la structure hiérarchique des phrases. Les pauses longues sont précédées par des voyelles plus longues et des variations plus grandes de l'intonation (Scott, 1982 ; Fernald et Simon, 1984 ; en Anglais et en Japonais : Fisher et Tokura, 1996, p. 348-349) [49] .

Quels sont les indices employés pour indiquer les frontières entre des segments linguistiques plus petits comme les mots ?

IV.1.2.La segmentation du signal de parole en mots

Les mots nouveaux sont souvent positionnés aux frontières prosodiques sur des pics de

[49] En musique, l'allongement des notes et la déclinaison de la hauteur se retrouvent avant les frontières musicales (Jusczyk et Krumhansl, 1993).

hauteur. Ces mots cibles sont mis en valeur dans des phrases courtes, avec une F0 plus haute, des variations de F0 plus importantes (Fisher et Tokura, 1996). Les nouveau-nés segmentent des mots comme « cup » dès l'âge de 7 mois et demi (Jusczyk et Aslin, 1995).

En Anglais, la segmentation se fait grâce à la différence entre les syllabes accentuées (stressed) et celles non accentuées (weak) (Cutler et Carter, 1987) [50] . En Français, la segmentation est syllabique, en Japonais, elle est fondée sur la mora. La langue maternelle influe sur le type de procédé de segmentation. Une fois ce processus déterminé, il fait alors partie du repertoire des mécanismes de traitement et sera utilisé quelle que soit la langue entendue. En effet, un français recourra toujours à une segmentation syllabique. Ainsi, la méthode de segmentation est originaire d'une première expérience d'apprentissage (Cutler, 1996).

Les bébés entre 6 et 9 mois sont sensibles à l'accent prédominant. Une stratégie raisonnable pour les enfants anglais consiste donc à placer le début de chaque mot sur chaque syllabe accentuée (Johnson et Jusczyk, 2001). Les syllabes de la parole se distinguent généralement par un pic d'amplitude qui est précédé et suivi d'une vallée d'amplitude (Jusczyk, 1997). A sept mois et demi, les enfants segmentent des mots avec un motif (fort – faible). En outre, dans ce dernier cas, si une syllabe faible suit la syllabe forte, la segmentation est effectuée (ex. : « guitar is » devient une seul unité : « taris »). Certains mots trisyllabiques sont correctement segmentés, lorsqu'ils ont un motifs « Fort, faible, Fort » (par exemple « parachute », au lieu de « para » et « chute » ; Houston et coll., 2000).

Dans l'expérience proposée par Saffran, Aslin et Newport (1996), les nouveau-nés sont testés avec des mots (par exemple, golatu, daropi) qui composent aléatoirment un signal de parole d'un langage artificiel, et des non-mots (part-word : tudaro, pigola) qui sont issus d'un mauvais découpage du signal de parole. Seuls des indices phonotactiques permettent de segmenter correctement les mots. Les nouveau-nés préfèrent alors écouter les mots isolés qui correspondent au signal de parole, prouvant ainsi qu'ils segmentent correctement le signal de parole. Johnson et Jusczyk (2001) ont repris cette expérience, proposée initialement avec une voie synthétisée, mais cette fois-ci avec une voie naturelle.

Johnson et Jusczyk (2001) ont alors étendu l'expérience de Saffran et coll. (1996), pour confronter l'influence de la phonotactique face aux signaux prosodiques. La première syllabe des parties de mots (part-word) est accentuée, ce qui suggère une frontière différente de celles indiquées par les statistiques. Ils ont trouvé que les enfants utilisent de préférence les indices prosodiques (dans ce cas l'accent). A l'âge de 10 mois et demi, les bébés utilisent donc des indices multiples (accents et phonotactique) pour la segmentation.

Est-il possible de traiter automatiquement ces indices, de façon à pouvoir les placer

[50] C'est généralement le cas dans les langues germaniques ainsi qu'en Russe, en Hongrois... En français, en revanche, c'est la dernière syllabe du mot qui est la plus accentuée ; en Espagnol et en Italien, la plupart des mots portent l'accent sur l'avant-dernière syllabe.

dans le discours ?

IV.1.3.Prédiction automatique des frontières prosodiques

Les frontières prosodiques contribuent à l'intelligibilité et à l'aspect naturel du discours. Ostendorf et Veilleux (1994) ont proposé un système de prédiction des frontières prosodiques à partir du texte, avec un minimum d'information syntaxique. Leur modèle hiérarchique prédit les ruptures entre différentes phrases prosodiques (intonative et intermédiaire) à partir du texte. L'apprentissage est automatisé et peut refléter des styles propres à différents locuteurs. Leur algorithme donne de bonnes performances en faisant appel à une description de la syntaxe réduite à plusieurs classes déterminées par une table. Les auteurs souhaitent prochainement valider leur travail par des jugements perceptuels des placements des ruptures prosodiques à partir d'une synthèse vocale.

IV.2.L'acquisition de la syntaxe

La prosodie intervient également dans l'enchaînement des mots entre eux, en particulier pour les structures syntaxiques, avec la désambiguïsation automatique des énoncés, mais elle peut aussi avoir un impact sur la détermination de certaines catégories syntaxiques

IV.2.1.Les structures syntaxiques

Venditti et coll. (1996) dressent une étude des rapports entre la syntaxe et la prosodie (intonation) pour le Japonais, le Coréen et l'Anglais. L'Anglais se distingue des deux autres langues par l'utilisation d'accents qui soulignent les mots importants, même si les tons coréens peuvent être interprétés comme des accents. La fréquence fondamentale est un marqueur de la syntaxe, mais aucun algorithme de prédiction ne peut être trouvé. Il n'est par conséquent pas possible de déterminer si la prosodie influe sur l'acquisition de la syntaxe, ou le contraire.

Cependant, les bébés sont sensibles aux rapports entre ces deux structures. Les enfants en bas âges préféreront entendre des paroles segmentées de façon à respecter la syntaxe. Ainsi, à l'âge de 9 mois, les bébés préfèrent le discours dont le sujet est séparé du verbe, tandis qu'ils ignorent celui dont la coupure est effectuée au centre de la phrase (Jusczyk et Kemler Nelson, 1996 ; Gerken, Jusczyk et Mandel, 1994).

La structure prosodique autour des pronoms et des noms sujets est différente. Une pause est encore insérée entre le sujet et le verbe, mais elle est artificielle dans le cas d'un pronom. Les bébés sont plus sensibles à la séparation marquée entre un nom sujet et le verbe (Jusczyk et Kemler Nelson, 1996). Les nourrissons perçoivent une frontière après un pronom dans une question inversée (« Do you like to play baseball » ? ; Gerken et coll., 1994). Les nourrissons de 7 à 10 mois préfèrent les pauses correspondant aux propositions uniquement dans le cas du discours adressé à l'enfant (Kemler Nelson, Hirsh-Pasek, Jusczyk et Wright, 1989). Dans ce contexte, la prosodie apparaît comme un des marqueurs possibles de la syntaxe des phrases. Mais, il existe seulement une congruence partielle entre la syntaxe et la prosodie. Effectivement, des contraintes liées à

la compréhension et à la production du discours peuvent imposer une régularité au niveau des phrases prosodiques (Ostendorf et Veilleux, 1994).

IV.2.2.Désambiguïsation de structures syntaxiques

Les indices prosodiques participent à la désambiguïsation syntaxique. Veilleux, Ostendorf et Wightman (1992) ont introduit un système de désambiguïsation entre deux découpages syntaxiques pour une même phrase. Cet article utilise deux composantes : une détection des ruptures et un algorithme calculant les relations entre ces ruptures et la syntaxe. Ces deux algorithmes sont implémentés par des arbres de décisions binaires. Ces relations entre syntaxe et prosodie sont apprises à partir d'une transcription écrite des énoncés. La détection automatique des pauses assigne un indice de ruptures à la fin des mots, en fonction des indices précédents et d'information acoustiques telles que la durée, ou l'estimation de la présence de tons. Elle s'effectue à partir du signal, annoté par des phonèmes. Les marqueurs indiquant la fin des mots sont trouvés automatiquement. Un arbre de décision classe les mots qui sont encodés par des vecteurs décrivant leur fin à partir des paramètres suivants : durée normalisée de la rime de la dernière syllabe, différence des durées entre l'attaque et la rime de la dernière syllabe, durée absolue de la pause, probabilité d'un ton de frontière, mesure locale du changement de vitesse du locuteur, et un marqueur indiquant la présence d'un accent lexical.

Ce système donne donc une séquence différente d'indices de ruptures (un nombre entier compris entre 0 pour les groupes clitiques à 6 pour une pause entre deux phrases) pour les deux interprétations parlées possibles d'une même phrase écrite. Les êtres humains obtiennent les meilleures performances lors de la discrimination de deux phrases parlées. Toutefois, les résultats de leur algorithme restent comparables aux performances humaines lorsque les indices de ruptures sont annotés manuellement. Effectivement, la détection automatique des ruptures diminue le score du système global. Les auteurs suggèrent d'améliorer cette composante en tenant compte d'autres indices prosodiques tels que la proéminence.

IV.2.3.Les catégories syntaxiques

Les informations phonologiques et prosodiques peuvent aider à distinguer les noms des verbes. Les adultes perçoivent le rythme trochaïque des noms dissyllabiques anglais, alors que les verbes ont un rythme ïambique (accent sur la seconde syllabe; Kelly, 1988). Les mots dissyllabiques qui peuvent appartenir aux deux catégories grammaticales mais qui ne sont pas différenciés par la position de l'accent, se distinguent par la durée de leurs syllabes et des différences dans leurs amplitudes (Sereno et Jongman, 1995). Les informations phonologiques, incluant les accents, la qualité des voyelles et la durée permettent de distinguer les mots de fonction (déterminant, préposition) des mots de contenu (nom, verbe, adjectif et adverbe) (Cutler, 1993 ; Gleitman et Wanner, 1982 ; Morgan, Shi et Allopenna, 1996 ; Cf. Chapitre 5).

IV.3.Données spectrales

Le champ de la prosodie peut s'étendre à l'étude du spectre (i.e. pour toutes les fréquences). Celui-ci apparaît comme une réalisation particulière du timbre à un instant précis. Des phénomènes spectraux ont lieu à mi-chemin entre la prosodie et la phonologie : la réalisation acoustique des voyelles, et le phénomène de coarticulation.

IV.3.1.Différence de réalisations des voyelles

Le discours adressé aux enfants peut contenir des indices acoustiques facilitant l'acquisition du langage. Ainsi, l'espace vocalique est augmenté dans le cas du discours vers l'enfant (Kuhl et coll., 1997). Pour trois langues (Russe, Suédois et Américain) le triangle vocalique [51] est plus étendu lorsqu'il s'agit du discours adressé aux enfants. Ainsi les deux premiers formants de chaque voyelle n'ont pas le même comportement, selon qu'il faut l'augmenter ou le diminuer pour augmenter le triangle vocalique. La distance acoustique entre chaque voyelle est plus élevée ce qui permet de mieux les distinguer (Kuhl et coll., 1992). Le triangle vocalique s'élargit pour les mots de contenu, lorsque le locuteur s'adresse à un enfant. Pour les mots de fonction, le contraire est vérifié. Le triangle est plus étendu pour les mots de fonctions lorsque le discours est adressé aux adultes [52] (Van de Weijer, 2001).

IV.3.2.Coarticulation

La coarticulation [53] est plus prononcée à l'intérieur des mots qu'entre les mots. Elle faciliterait alors à la fois la segmentation et la reconnaissance des mots. La coarticulation dépend en partie de l'intervalle entre eux. Par exemple, la coarticulation effectuée sur le phonème [k] dans « coo » est moins importante que dans « clue », où [k] est séparée de la voyelle par [l]. Dans « sack Lou », la coarticulation est encore moins marquée (Ladefoged, 1975). Ainsi la coarticulation est influencée par les frontières des mots. En outre, les signatures prosodiques contiennent un renforcement articulatoire des consonnes aux jonctions prosodiques. A l'âge de 5 mois, les bébés sont sensibles à certaines variations coarticulatoires dans le discours. Effectivement, ils utilisent préférentiellement la coarticulation aux indices phonotactiques afin de segmenter le signal de parole en mot (Johnson et Jusczyk, 2001).

IV.4.Perception des indices locaux

Les indices prosodiques doivent pouvoir être situés dans les unités linguistiques comme la syllabe ou le mot pour la maîtrise du langage. Une expérience ancienne

[51] Le triangle vocalique est l'espace désigné par 3 points donnés par les coordonnées des deux premiers formants (F1, F2) pour les trois voyelles /i/ /a/, /u/. Les voyelles des mots « sheep » et « shoes » ont été comparées chez dix mères. Dans le cas du « parler bébé », la F0 est plus élevée, F1 est le même que chez l'adulte, et F2 est significativement plus élevé pour /i/ et plus bas pour /u/ (Andruski & Kuhl, 1996).

[52] Des trois locuteurs étudiés, tous ne présentent pas l'expansion de l'espace vocalique.

[53] La surimposition des articulations adjacentes d'après Ladefoged (1975).

(Ladefoged, 1975) montre qu'un clic placé à l'intérieur d'un phonème ne semble pas avoir une localisation précise. L'impossibilité de localiser ces clics à l'intérieur des séquences phonétiques formant des phrases s'accompagne d'une incapacité à localiser directement des phonèmes. Cependant, ces clics sont émis par une source différente du signal de parole, contrairement aux informations prosodiques. La localisation des pics intonatifs est peut-être rendue possible, parce que cette dimension est intiment liée au contenu spectral.

Les nourrissons américains âgés de un à deux mois sont capables de faire la différence entre des stimuli multisyllabiques qui se distinguent uniquement par la position de l'accent (Spring et Dale, 1977). Ces résultats ont été étayés en 1993, en montrant que les nourrissons sont sensibles aux différences entre le rythme trochaïque et ïambique (Jusczyk, Cutler et Redanz, 1993). A 9 mois, les bébés anglais préfèrent les mots anglais accentués sur la première syllabe et ne prêtent pas attention aux syllabes non accentuées (Jusczyk et coll., 1993). Cette négligence ne s'explique pas par un phénomène articulatoire ou perceptuel. Les premiers mots sont bien formés, mais contiennent seulement l'information prosodique que les bébés sont en mesure de saisir (Demuth, 1996).

Pour les adultes, les voyelles « hyper articulées » sont considérées comme étant plus représentatives de leurs classes. Pour parler, les enfants doivent reproduire les sons qu'ils entendent dans leur propre domaine de fréquence. Ainsi les représentations de la parole seraient encodées en mémoire dans des dimensions spectrales abstraites. Dans ces conditions les enfants peuvent catégoriser des unités phonétiques parlées par des locuteurs différents. Cette représentation leur permet également de définir une mesure indépendante de la fréquence, qui leur permet de produire des unités équivalentes avec leur conduit vocal (Kuhl, 1991).

Les mécanismes permettant le traitement des indices linguistiques rapides sont mal connus. Les enfants dysphasiques (SLI : Specific Language Impairment) ont des difficultés pour maîtriser le langage, et ces problèmes pourraient avoir pour origine un déficit du traitement des indices rapides. Ces enfants ne peuvent couramment pas identifier des stimuli rapides à l'intérieur de la parole. Des durées inférieures à quelques dizaines de millisecondes ne peuvent être traitées par ces enfants. Or, cette durée correspond justement à de nombreux contrastes phonétiques. Ce déficit est désigné par les termes d'un trouble du traitement auditif temporel (Benasich et Tallal, 2002) [54] .

Des études utilisant l'écoute dichotique (Schwartz et Tallal, 1980), les enregistrements intracorticaux (Liégeois-chauvel, De Graaf, Laguitton et Hauvel, 1999) et l'imagerie (Belin et coll., 1998 ; Fiez et coll., 1995) ont suggéré que l'hémisphère gauche répond préférentiellement aux motifs acoustiques changeant rapidement. C'est pourquoi le traitement phonologique serait latéralisé. Fitch, Brown, O'Connor et Tallal. (1993) ont retrouvé ces résultats chez le rat, pour des séquences rapides de tons. Liégeois-chauvel et coll. (1999) ont montré que les aires centrales gauches encode le VOT (voice-onset time) d'une consonne [55] .

[54] Nous reviendrons sur cette hypothèse pour l'acquisition de la syntaxe (voir Chapitre Six).

Zatorre et Belin (2001) ont étudié les mécanismes employés par le cerveau pour traiter les durées ou l'information spectrale à l'aide d'une TEP [56] . Le cortex auditif central des deux hémisphères répond aux variations temporelles, alors que les aires temporales antérieures supérieures répondent aux variations de fréquences. Lorsque les tons se succèdent très rapidement, une lésion située à gauche provoque plus de dégât qu'une lésion à droite (Samson Ehrlé et Baulac, 2001).

Ainsi il y aurait deux systèmes de traitement, l'un réservé aux différences spectrales (plus particulièrement pour la tonalité), l'autre pour traiter les changements rapides (Zatorre et Belin, 2001). D'autre part, ces différents mécanismes n'auraient pas les mêmes bases biologiques. L'organisation ainsi que la nature des neurones de l'hémisphère droit permettraient un traitement plus rapide.

V.Conclusion

Cet aperçu des propriétés prosodiques reprend l'ordre dans lequel seront abordés les différents thèmes de cette thèse. Ce bilan s'appuie sur des domaines différents qui constituent un Continuum Temporel. Ainsi, l'identification des langues s'appuie sur les données suprasegmentales [57] . Ensuite, l'identification des attitudes prosodiques sera abordé au travers de l'identification des contours intonatifs [58] . Enfin, l'identification des mots de fonction et de contenu portera sur l'utilisation d'indices prosodiques, définis localement. Ce dernier point sera également examiné dans le contexte particulier d'un déficit pour le traitement auditif rapide, pour fournir une modélisation du processus réalisé par les enfants SLI. Ainsi, nous prouverons les capacités du réseau TRN pour le traitement des structures temporelles de la parole, en appuyant notre travail sur des tâches diverses du traitement de la prosodie, réalisées à partir de données acoustiques.

[55] Cette durée de voisement au début d'un phonème (VOT : voice onset time) distingue les syllabes /ba/ et /pa/ (Liberman, 1996).

[56] Des séquences de tons purs qui varient suivant deux dimensions différentes : soit suivant leur durée (seulement deux fréquences possibles), soit suivant leur fréquence fondamentale (seulement une durée possible). En diminuant la durée des stimuli, l'hémisphère gauche est plus sollicité, et en augmentant le nombre de fréquence, l'hémisphère droit est plus activé.

[57] Effectivement, les informations locales n'ont pas reçu une attention particulière dans notre travail, même si elles revêtent une grande importance pour ce problème particulier.

[58] Dans notre travail, nous ne considérons pas les autres informations prosodiques, même si elles sont utiles à la réalisation des attitudes étudiées.

Chapitre Trois Thème 1 : Identification Automatique des Langues (I.A.L.)

« The Babel Fish is small, yellow and leech-like, and probably the oddest thing in the Universe. It feeds on brainwave energy received not from its own carrier but from those around it. The practical upshot of all this is that if you stick a Babel Fish in your ear you can instantly understand anything said to you in any form of language. The speech patterns you actually hear decode the brainwave matrix which has been fed into your mind by your Babel Fish. Meanwhile, the poor Babel Fish, by effectively removing all barriers to communication between different races and cultures, has caused more and bloodier wars than anything else in the history of creation. » Douglas Adams, The Hitchhiker's Guide to the Galaxy (Harmony Books, 1979).

I.Quelques notions sur l'Identification Automatique des Langues

Le premier thème que nous abordons est l'Identification Automatique des Langues. Notre objectif est d'appliquer le réseau TRN au traitement de séquences acoustiques de longues durées. Ainsi, celui-ci devra montrer ces capacités pour identifier une langue en

tenant compte de la contrainte temporelle. Dans ce cas, le réseau peut donner une définition globale du rythme d'une langue, ce qui validera le premier point de l'hypothèse de Continuum Temporel. Commençons par résumer les travaux concernant l'IAL.

I.1.Définition

Si le mythe de la tour de Babel peut être réalisé, la langue parlée devra être devinée à l'écoute d'un passage de parole énoncé par un locuteur inconnu (Muthusamy, Jain et Cole, 1994). Cette définition simple cache un axe de recherche complexe : L'Identification Automatique des Langues (IAL). La première motivation de l'IAL est d'utiliser des instruments qui puissent traiter le signal de parole le plus directement possible, sans utiliser des ressources linguistiques trop importantes et complexes. Les phrases doivent être analysées dans leur ensemble afin d'en extraire une signature de chaque langue. Cette tâche est ardue, puisque le signal de parole contient de multiples informations comme le sexe, l'âge, les accents, le statut émotionnel, qui s'ajoutent à la langue parlée. En outre, le signal acoustique est souvent altéré par des bruits dus à l'environnement, ou au canal de communication.

L'IAL, un domaine de recherche plutôt récent, devrait tirer parti du plus grand nombre possible de domaines d'expertise. Ces recherches s'appuient sur l'ingénierie du Traitement Automatique de la Parole, mais elle est maintenant rejointe par la psycholinguistique, et bien d'autres disciplines devraient suivre. La problématique de l'IAL se comprend mieux en évoquant un exemple concret. Les urgences américaines (911) gèrent 140 langues, avec des interprètes humains experts. On relève un délai de 3 minutes pour que la langue Tamil soit identifiée. L'interlocuteur, en effet, a été d'abord dirigé sans succès vers trois interprètes du Sud-est asiatique, et a écouté plusieurs enregistrements de remerciements dans d'autres langues. Le délai aurait encore été rallongé s'il n'avait pas prononcé le mot « Tamil » en Anglais (Muthusamy et coll., 1994).

I.2.Les enjeux

Nous sommes à une époque de **communication multilingue**, que ce soit entre êtres humains (au sein des grandes mégalopoles ou par téléphone interposé), ou entre humains et machines (domaine des Interfaces Homme-Machine ou **IHM**). A cela s'ajoute une demande croissante pour des applications de traitement automatique de la parole qui interviennent de plus en plus dans notre quotidien (dictée automatique, lecture de messages électroniques à distance,…).

Ce constat implique le développement d'applications capables de gérer plusieurs langues et/ou d'identifier une langue parmi d'autres. De tels systèmes peuvent être envisagés dans des tâches d'assistance au dialogue humain (réservation dans des hôtels), au sein d'IHM ou encore dans le cadre de l'indexation multimédia basée sur le contenu, un secteur en pleine expansion. Un système d'IAL prendrait également sa place en amont d'un système automatique de traduction. Le nombre de langues à traduire est très élevé jusqu'à 120 langues différentes, pour une entreprise de traduction.

L'objectif d'un logiciel d'IAL est de pouvoir diriger un locuteur avec un système de traitement de la parole (humain ou automatique) adapté à sa langue. Un tel système prendrait sa place dans le domaine militaire, en identifiant la langue d'un sujet étranger. Une application prévue pour l'IAL pourrait facilement être adaptée à l'enseignement des langues. Ainsi, un système d'IAL s'appuyant sur la prosodie pourrait corriger l'intonation. Evidemment, l'IAL peut être étendue à la reconnaissance des accents régionaux, ainsi que des dialectes.

A ces enjeux techniques s'ajoutent nombre de motivations scientifiques, aussi bien d'un point de vue linguistique que sur le plan de la cognition. L'ontogenèse des représentations cognitives des langues est en effet un mécanisme encore mal connu, et plusieurs travaux récents attirent l'attention sur le rôle joué par la prosodie et le rythme dans ce processus (Nazzi et coll., 1998 ; Ramus et coll., 1999 ; Ramus, 2002a). Un système d'IAL permettrait également de proposer une topologie complémentaire des langues, en les classant d'après leurs ressemblances et leurs dissemblances.

I.3.Objectifs et plan

Avant d'aborder les expériences menées en Identification Automatique des Langues, un bilan des techniques utilisées sera proposé, avant de décrire les expériences perceptuelles. Cette partie bibliographique se terminera par les travaux adressant le traitement de la prosodie dans les applications d' l'IAL.

Pour la partie expérimentale, nous aurons pour premier objectif de montrer que le réseau TRN peut contribuer à l'identification automatique des langues, en étudiant deux types de représentations du signal de parole :

1. Une description en consonnes et voyelles, rendant compte uniquement de la dimension rythmique ;
2. Une représentation spectrographique, mêlant plusieurs dimensions acoustiques.

Notre second objectif sera de prouver que le rythme d'une langue influence sa reconnaissance avec le réseau TRN, de la même façon que chez les nourrissons (Nazzi et coll., 1998). Dans ce but, nous proposerons une simulation de la distinction des classes rythmiques à partir du signal de parole compris dans les basses fréquences du signal.

Ces travaux ont été conduits avec les corpora LSCP, MULTEXT et OGI-MLTS accessibles à la communauté scientifique.

II.Contexte de l'IAL

Cette partie aborde le thème de l'IAL sous trois angles :

· technique, pour une description des systèmes existants et les composants de la

parole, qu'ils utilisent.

- perceptuel, où les expériences conduites avec des êtres humains (bébés et adultes) et avec des primates seront présentées.
- prosodique, avec un examen des parutions traitant de l'IAL à partir de la prosodie, cadre dans lequel s'inscrit notre propre étude.

II.1.Les bases techniques d'un système d'IAL

II.1.1.La jeunesse de l'IAL

Les premières publications [59] ayant trait à l'IAL ont vu le jour dans les années 70-80, pour la recherche militaire (Dutat, 2000). Ces publications fournissent très peu de renseignements sur le corpus employé et n'apportent pas d'indications particulières comme les conditions d'enregistrement, le nombre de locuteurs, leur sexe, etc.

Beaucoup d'idées avaient été examinées, mais elles ont eu peu d'impact sur les recherches actuelles. Les premières études étaient basées sur les différences spectrales entre les langues. Le spectre pouvait être représenté de plusieurs manières : par des motifs (features vectors), à partir des formants, ou encore par l'utilisation des coefficients sceptraux (Zissman et Berkling, 2001). La similarité spectrale est ensuite déterminée par une distance euclidienne, ou de Mahalanobis, ou par le biais de l'algorithme de regroupement des k-moyens. Elle est calculée entre le vecteur de test et tous ceux de la base d'apprentissage, la distance minimum indique la langue identifiée. Cette solution a été remplacée par une mixture gaussienne dans les travaux les plus récents [60] . Sur quelle structure est fondé un système d'IAL ?

II.1.2.Les architectures d'un modèle d'IAL

Un système d'IAL se construit toujours en deux étapes : apprentissage et validation. Ainsi, l'étape d'apprentissage établit un modèle de chaque langue à reconnaître, à partir d'un ensemble de données, constituées du signal de parole de plusieurs locuteurs, de segmentation phonémique, de motifs articulatoires. En phase de test (validation ou développement [61]), le signal de parole à identifier est comparé aux modèles de langues construits, afin de déterminer la langue de l'extrait (Figure 3.1). L'une des difficultés majeures de l'IAL est de bien justifier que c'est la langue parlée elle-même qui est

[59] 14 publications auraient vu le jour avant 1993 essentiellement dans des congrès et colloques comme Eurospeech, ICASSP, SRS (Muthusamy et coll., 1993).

[60] D'autres techniques d'apprentissage comme les systèmes experts, les algorithmes de regroupement (clustering), les classifieurs quadratiques ont été testées (Zissman, 2001 ; Muthusamy et coll., 1993).

[61] Le corpus OGI-MLTS dresse une liste des fichiers répartis en trois corpus : apprentissage, développement et validation. La validation est la phase du test final. Le développement est une étape intermédiaire, qui permet de vérifier la validé des indices retenus pendant l'apprentissage.

reconnue, et non le locuteur (ou d'autres indices présents dans le signal enregistré). Ainsi, pour éviter que le système identifie le locuteur, les bases d'apprentissage et de test contiennent toujours des locuteurs distincts (Dutat, 2000).

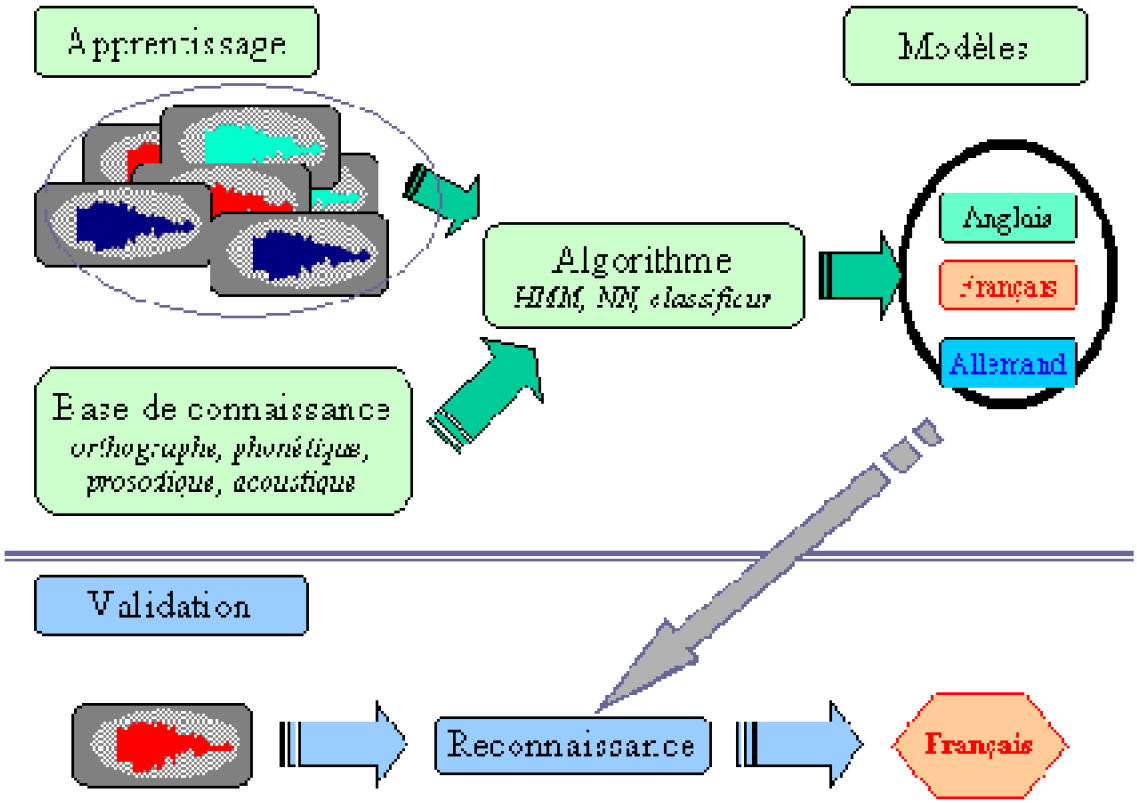


Figure 3.1 Schéma de l'algorithme d'apprentissage en IAL.

Avant d'effectuer l'identification, les signaux de paroles sont le plus souvent traduits sous la forme d'unités symboliques. Deux solutions sont possibles pour assigner une valeur discrète à un motif acoustique (par exemple, lors de la reconnaissance des phonèmes) :

- Un modèle est établi pour l'ensemble des langues (Zissman et Berkling, 2001). Chaque signal de parole est traduit dans une représentation unique indépendante des langues à identifier.
- L'encodage en unité discrète est dépendant des langues à reconnaître. Dans ce cas, le passage à identifier est encodé dans chaque langue. Concrètement, le nombre d'encodeurs est limité au nombre de langues ayant les données nécessaires à la fabrication d'un système de reconnaissance de phonèmes. Cette technique est beaucoup plus coûteuse que la précédente, tant d'un point de vue informatique qu'humain, lors de la segmentation en unité linguistique (Figure 3.2).

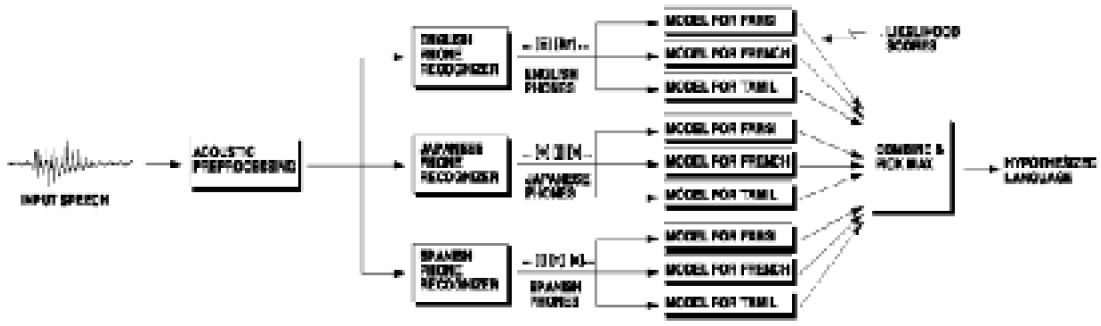


Figure 3.2 Un système basé sur un ensemble de systèmes de reconnaissance de phonèmes (tiré de Zissman et Berkling, 2001).

Ainsi, les applications d'IAL se subdivisent en deux types: les systèmes non supervisés où les motifs acoustiques permettant la distinction des langues doivent être trouvés de manière automatique et les systèmes supervisés pour lesquels les motifs acoustiques sont appris. Les systèmes non supervisés sont généralement plus efficaces pour les courtes durées et les systèmes supervisés donnent les performances les plus élevées en IAL [62] (Zissman et Berkling, 2001).

II.1.3.Systèmes supervisés

L'approche la plus classique est constituée d'un système de reconnaissance de phonèmes, qui transforme le signal de parole en suite d'éléments discrets. Enormément de progrès ont été effectués dans le traitement de la parole indépendamment du locuteur à partir de méthodes comme les Chaînes de Markov Cachées (HMM) ou les réseaux de neurones artificiels. Ces méthodes d'apprentissage permettent de modéliser les phonèmes en fonction de leur contexte. Cet apprentissage s'effectue sur un ensemble de signaux de paroles, pré-segmentés et étiquetés en phonèmes par des experts phonéticiens. Un modèle de la langue est alors établi par un traitement statistique de leur distribution.

Ces systèmes incluent donc des connaissances a priori sur le signal de parole. Ils sont donc coûteux, puisqu'ils nécessitent un traitement humain qui permet de localiser les phonèmes dans le signal de parole. Effectivement cette étape est nécessaire pour effectuer l'apprentissage des systèmes reconnaissant les phonèmes, même si ils ne sont pas utilisés lors de la reconnaissance de la langue pendant la phase de validation.

Il n'existe actuellement pas assez de données pour permettre l'utilisation d'une reconnaissance de phonèmes pour toutes les langues. Seulement 6 langues du corpus OGI-MLTS sont étiquetées. Cependant, il est possible de se limiter à celles-ci, en ne travaillant qu'avec 6 systèmes de reconnaissance de phonèmes et d'utiliser 11 modèles phonotactiques, pour chaque modèle acoustique, soit 6x11 modèles phonotactiques (Muthusamy et coll., 1993).

[62] Les thèses de F. Pellegrino (1998) et H.P. Combrinck (1999) dressent des tableaux des performances de ces deux types de systèmes.

Muthusamy et coll. (1993) ont comparé des méthodes basées sur des composantes acoustiques (70 %), des catégories phonétiques larges (83.2 %) et une classification phonétique plus fine (86.3 %) sur une tâche de discrimination de l'Anglais et du Japonais, extraits du corpus OGI-MLTS. Des classes phonétiques détaillées permettent une identification supérieure à celles obtenues avec des catégories phonétiques plus larges, même si ces dernières contiennent moins d'erreurs lors du décodage phonétique.

Dans le cadre de l'identification des langues européennes (Anglais, Français, Espagnol, Portugais, Allemand et Italien), Caseiro arrive à des scores de 79 % (pour 5 secondes de signal de parole), en utilisant un seul système de reconnaissance des phonèmes du portugais (Chaîne de Markov Cachées). Il faut pourtant recourir à un dictionnaire de prononciation et une transcription orthographique (Caseiro et Transcoso, 1998). La même optique est employée pour l'identification de 10 langues du corpus OGI-MLTS avec 10s de signal (59.7 % ; Lamel et Gauvain, 1994).

Une implémentation unifiée des dimensions phonotactique, acoustique, phonétique et prosodique est décrite par Hazen et Zue (1994 ; 1997). Le système comprend trois étapes : 1) un pré-traitement qui permet d'obtenir une description acoustique du signal [63] , 2) une reconnaissance phonétique à partir du pré-traitement, 3) un classifieur qui identifie la langue. La fréquence fondamentale est traduite par son logarithme et sa valeur moyenne lui est soustraite. La reconnaissance phonétique est assurée par l'algorithme SUMMIT avec 87 unités phonétiques indépendantes de la langue. La normalisation des canaux [64] et l'optimisation des paramètres induit une augmentation des performances. Ils atteignent 78.1 % d'identification pour 10 langues d'OGI-MLTS (45s de signal). Des système équivalents atteignent 88.8 % (Zissman et Singer, 1995) et 90.8 % pour Yan et Barnard (1995).

Récemment les performances de différents systèmes d'identification automatique des langues ont été comparées. Les systèmes les plus performants atteignent 10 % d'erreur pour 10 langues à identifier (Zissman et Berkling, 2001). Cependant, de tels résultats ne peuvent être obtenus qu'à partir de nombreuses heures de discours étiqueté, ce qui pose problème pour l'identification de nouvelles langues. Le recours à des systèmes non supervisés est un palliatif à ce défaut.

II.1.4.Systèmes non supervisés

Dans ce cas, les informations fournies au système ne sont pas étiquetées. Kwasny, Kalman, Wu et Engebretson (1992 ; cité dans Combrinck, 1999) ont utilisé des réseaux récurrents sur des fichiers de parole non traitée. Les résultats sont excellents, mais le corpus utilisé est restreint (2 locuteurs, 2 phrases pour 2 langues). Du Preez et Weber (1998) ont reporté des résultats remarquables pour des chaînes de Markov cachées de rang supérieur (« high order HMMs »). A l'image des réseaux récurrents, une image de

[63] Les signaux de paroles sont traduits par 14 coefficients MFCC toutes les 5 ms, avec une fenêtre de hamming de 25.6ms, et une transformée de Fourrier discrète. Les variations de chaque canal sont également prises en compte (14 coefficients delta-MFCC).

[64] Une normalisation aveugle des canaux est effectuée : pour chaque phrase, la valeur moyenne de chaque canal est calculée et soustraite à chacune des valeurs individuels.

l'activité passée est maintenue. Les expériences ont été conduites sur l'Anglais et l'Hindi du corpus OGI-MLTS, et ils atteignent 79.8 % en 5s, et 97.4 % pour 45s (Combrinck, 1999).

Zissman (1993, 1996) emploie des chaînes de Markov cachés et des mixtures gaussiennes, poursuivant les travaux de Nakagawa et coll. (1992). Chaque langue est représentée par un ensemble de distributions gaussiennes multivariées, défini pendant la phase d'apprentissage. Une phrase test est classée en cherchant le modèle qui lui ressemble le plus (Zissman, 1996 : 50 % pour 5s de signal pour 10 langues du corpus OGI-MLTS)

II.1.5.Domaine linguistique différenciant les langues

Les différences observées entre les langues peuvent s'organiser en quatre catégories linguistiques (Zissman et Berkling, 2001) :

- La **phonologie** étudie les phonèmes, une unité phonologique caractéristique d'une ou plusieurs langues. Ils peuvent être traduits de plusieurs façons acoustiques suivant la langue, l'accent et le locuteur lui-même. Une langue se distingue par le répertoire des phonèmes qu'elle emploie, mais aussi par la façon dont ces unités s'enchaînent au sein du discours (phonotactique). Cependant, il arrive souvent que des phonèmes ou leur combinaison se retrouvent dans plusieurs langues. Dans ce cas, la fréquence d'occurrence de ces motifs permet de distinguer ces langues. Nous venons de voir que cette particularité est le plus souvent utilisée par les méthodes développées en IAL.
- La **morphologie** étudie la manière dont sont formés les mots, ainsi que le lexique qui permettent de caractériser une langue.
- La **syntaxe** définit l'organisation des mots entre eux.
- La **prosodie** possède des données acoustiques différentes pour chaque langue (durée, tonalité, accents, etc.).

Nous avons donné précédemment un bref aperçu des techniques et des performances en IAL. L'un des attraits de l'IAL est de recruter plusieurs composantes de la linguistique. Cependant, ces différents domaines ne sont pas encore tous intégrés de façon optimum dans un système d'IAL. Des expérimentations perceptuelles ont été menées afin de déterminer les indices utilisés par les adultes, les enfants et également les primates, lors de l'identification d'une langue.

II.2.Etudes perceptuelles

Cette section est consacrée aux expériences menées chez les primates humains et non-humains pour étudier leurs capacités à distinguer ou identifier des langues. Deux types d'études peuvent être spécifiées :

1. l'identification à partir de parole naturelle, afin de répertorier les indices perçus.

2. la discrimination de langues à partir de parole modifiée par resynthèse ou par filtrage, afin de s'assurer de la validité des indices mentionnés dans les études précédentes, en éliminant les autres d'indices disponibles.

Pour chacun de ces cas, nous examinerons les réponses obtenues par les adultes, les nourrissons et également les primates non-humains.

II.2.1.Parole naturelle

II.2.1.1.Adultes

Les études traitant des expériences perceptuelles en IAL sont moins nombreuses que les publications sur les systèmes eux-même. 5 études seulement ont été effectuées avec des sujets humains adultes : Muthusamy et coll., 1994 ; Stockmal, Muljani et Bond, 1996 ; Bond, Stockmal et Muljani, 1998 ; Lorch et Meara, 1989 et 1995).

Avant de présenter un système d'IAL en 1996, Muthusamy et coll. (1994) proposent d'étudier les indices utilisés par les êtres humains [65] . Des connaissances dans une langue donnée facilitent son identification (taux d'identification de 44 % lorsqu'une seule langue est connue à 67 % pour quatre langues). Bond et Fokes (1991) avaient également montré que l'exposition à une langue est un facteur déterminant dans la réussite de son identification. Ainsi il est plus facile pour des Coréens d'identifier du Chinois ou du Mandarin, ou pour des Espagnols d'identifier du Sud-américain. Deux facteurs socio-linguistiques jouent un rôle primordial dans la reconnaissance des langues : la familiarité avec les langues et les particularités de la langue maternelle. La durée des extraits et le nombre de langues connues sont également des facteurs de la performance d'identification (Vasilescu, Hombert et Pellegrino, 2000).

Un second type d'étude doit donc être envisagé, pour déterminer comment des sujets humains se comportent face à une langue inconnue. Les sujets devaient juger si deux passages de langues étrangères correspondaient à la même langue, et expliquer les raisons de leur choix. Lorch et Meara (1995) testent la discrimination de deux langues étrangères (Farsi et Grec). Les score moyens sont relativement peu élevés 62.6 % au premier essai, et 64.9 % pour le suivant. Le meilleur score est de 88 %, et certains sujets répondent très en dessous du hasard, alors que l'identification des voix est aisée (96 %). La faiblesse de performances peut s'expliquer par la présence de sons existant en Anglais. A travers cette étude, la discrimination de deux langues inconnues semble relativement difficile pour des êtres humains.

Une expérience plus récente propose d'étudier si des adultes peuvent créer une représentation du Japonais afin de discriminer celui-ci de l'Arabe, du Russe, de l'Indonésien et du Chinois. (Bond et coll., 1998). Deux questions ont été abordées :

1. Est-ce qu'une exposition minimale à une langue peut améliorer son identification ?

[65] L'identification des langues est effectuée avec 10 sujets anglais et 2 sujets ayant pour langue maternelle chacune des neuf langues restantes. Les signaux de paroles sont tirés de la base OGI-MLTS. Les extraits sont issus des passages de parole spontanée (autour de 1 minute) et contiennent moins d'une moitié de silence.

Quelle est l'influence de cette exposition ? 2.

La première expérience concerne des expositions de courtes durées (liste de mots ou histoires), pour deux périodes de 5 ou 15 minutes. Seule l'histoire donne une idée de la prosodie d'une langue. Le matériel audio ne contient pas les mots les plus connus qui permettraient l'identification d'une langue. L'exposition de 5 minutes a un effet sur les scores d'identification. Les sujets créent une représentation stable du Japonais, mais elle n'est pas conservée en mémoire après une pause de 30 minutes. Comme l'exposition à un seul locuteur pourrait être la cause de cet échec [66] , l'expérience a été alors adaptée avec trois locuteurs différents. Lorsque le test d'identification suit immédiatement l'entraînement, les performances sont améliorées. Cependant, ces effets disparaissent toujours au cours du temps. En outre, quelques sujets ont des scores élevés, mais la majorité a des difficultés pour extraire les indices provenant de plusieurs locuteurs [67] .

La prosodie est la seule dimension accessible au sujet qui ne maîtrise pas la langue, mais d'autres indices sont présents dans le discours. Les auditeurs sont également gênés par l'arrivée massive d'information dans le cas d'un discours. Effectivement les sujets essayent de découvrir des mots dans le discours, travail qu'ils n'ont pas à effectuer dans le cas des mots isolés.

Il est difficile de savoir si les sujets séparent bien des langues distinctes et non des locuteurs distincts. Stockmal et Bond (1998) ont alors réalisé une expérience avec un seul locuteur parlant 2 langues, inconnues des sujets, qui ont cependant réussi à séparer les deux langues.

L'objectif principal de toutes ces études est d'isoler les indices utilisés par les sujets pour distinguer des langues. La complexité des systèmes vocaliques (Français, Portugais vs. Italien, Roumain, Espagnol), et la présence de segments consonantiques spécifiques (la présence des dentales fricatives /S Z/ en Roumain et Portugais) entrent en jeu lors de l'identification d'une langue (Vasilescu et coll., 2000). Muthusamy et coll. (1994) répertorient un certain nombre de distinctions sur des phonèmes caractéristiques de certaines langues :

- Son aspiré pour le Farsi, occurrence fréquente de /sh/
- Beaucoup de sons nasaux pour le Français
- Allemand : le mot ich est reconnu, « harch » son aspiré (vélaire) ; confusion avec le Farsi
- Japonais : « crips » stops, des mots distincts sont reconnus : watashiwa et mashita
- Coréen : le mot imnida.
- L'Espagnol est caractérisé par certains sons comme la paire « eh-s »

[66] Une étude a montré que la discrimination des phonèmes /r/ et /l/ par auditeurs Japonais –qui ne connaissent pas cette distinction phonémique- est facilitée lorsque les sujets entendent plusieurs locuteurs (Lively, Logan et Pisoni, 1993).

[67] Des études ont également montré que la variabilité des locuteurs affectait la mémoire lors de la rétention de liste mots (Goldinger, Pisoni et Logan, 1993 ; Martin, Mullennix, Pisoni, Summers et Palmeri, 1989 ; Palmeri, Goldinger et Pisoni, 1993).

- Tamil /r/ et /l/
- Le Vietnamien contient plusieurs nasales distinctives, vélaire nasale /ng/

Les sujets mentionnent également certaines dimensions prosodiques :

- Le **Timbre de la voix** oula qualité de la voix rappelle celle d'un de leur proche étranger [68] (Muthusamy et coll., 1994 ; Stockmal et coll., 1996 ; Lorch et Marea, 1995).
- L'**Intonation** pour les langues utilisant des excursions de l'intonation sur des courtes durées [69] . Ainsi, le Japonais et Chinois sont regroupés dans la représentation perceptive [70] proposée par Stockmal et coll. (1996). La même confusion entre le Vietnamien, le Chinois Mandarin et la Japonais est observée chez Muthusamy et coll. (1994), mais avec des locuteurs multilingues. Le Français (Muthusamy et coll., 1994) et l'Italien (Vasilescu et coll., 2000) sont également reconnus par leur intonation.
- Le **Rythme** : Les auditeurs se plaignent que la vitesse est trop rapide pour certaines langues étrangères (Stockmal et coll., 1996 ; Lorch et Marea, 1995). L'Espagnol donne l'impression d'un débit de parole élevé (Muthusamy et coll., 1994). Des sujets anglais semblent reconnaître des mots dans le cas de l'Arabe et du Russe (Stockmal et coll., 1996). Une segmentation en mot pourrait être effectuée par les sujets dans la mesure où l'Arabe et le Russe appartiennent aux langues accentuelles tout comme l'Anglais, langue maternelle des sujets de ces expériences. Ceci conforte l'idée que les différences rythmiques de ces langues pourraient être liées aux techniques de segmentation effectuées par les sujets (Cutler, 1996).

L'intégration de toutes les dimensions constitue la clef d'une distinction réussie entre plusieurs langues (Stockmal et coll., 1996). Les adultes utilisent plusieurs sources d'informations, tandis que les enfants scolarisés utilisent principalement des indices segmentaux (Bond et coll., 1998). Les plus jeunes sont influencés quant à eux par la voix du locuteur, et des indices prosodiques comme le rythme ou la tonalité [71] .

II.2.1.2.Nouveau-nés et nourrissons

Les premières expériences avec des nouveau-nés ont d'abord conduit à penser qu'ils différenciaient une langue familière d'une langue étrangère. Mehler et coll. (1988) ont montré que les nouveau-nés distinguaient le Russe du Français, et l'Anglais de l'Italien [72] .

Tableau 3.1 Discrimination des langues par les nourrissons [73]

68 Cette idée se rapproche des techniques d'identification du locuteur, mises en œuvre par Li et Edwards (1994) qui s'appuient sur les propriétés spectrales des voyelles pour l'AL.

69 Langues à tons comme le Chinois Mandarin.

70 Cette expérience est conduite avec des sujets Américains qui connaissent mal les langues asiatiques.

71 En citant un article de Burnham et Torteson (1995), Bond et coll. (1998) précisent que l'utilisation de la prosodie ne s'étend pas après 5 ans chez les enfants.

72 L'étude citée montre que les nouveau-nés ne distinguaient pas l'Italien du Français, Mehler et Christophe (1995) reviennent sur ces données, et prouvent finalement qu'ils distinguent ces deux langues.

73 L'astérisque indique les études basées sur la prosodie obtenue par un filtre passe-bas. Ramus (2002) utilise une procédure de resynthèse.

Age	Langues	Référence
4 mois	Anglais / Espagnol	Bahrick et Pickens, 1988
2 jours	Maternel / étranger	Moon et coll., 1993
4 jours et 2 mois	Maternel / étranger	Mehler et coll., 1988 *
4 jours	2 langues étrangères	Mehler et coll., 1988 ; Mehler et Christophe, 1995
2 mois	Seulement Maternel / étranger	
4 jours (bébé français)	Anglais / Japonais Mais pas Anglais/hollandais	Nazzi et coll., 1998 *
2 à 5 jours	Néerlandais / Japonais	Ramus, 2002 (Resynthèse)

Le tableau 3.1 énumère les principales études de discrimination de langues chez le nourrisson. Nous reviendrons dans la section II.2.2 sur celles employant de la parole modifiée.

II.2.1.3.Primates non-humains

Des singes tamarins ont été testés pour la discrimination Néerlandais-Japonais, à partir de la parole naturelle, et de la parole jouée à l'envers. La condition de parole naturelle a été effectuée avec succès par les singes, alors que les nouveau-nés échouaient à cause de la multiplicité des voix (Ramus, Hauser, Miller, Morris et Mehler, 2000). Si les primates non humains parviennent à effectuer cette tâche, les aptitudes de discrimination des langues se fonderaient donc sur des capacités de traitement auditif général, et non spécifiques au langage.

Les études perceptuelles ont pour objectif de mettre à jour les indices potentiels pour l'identification automatique des langues. Cependant, il reste encore à démontrer si ces indices sont bien utilisés lorsqu'ils sont présentés isolément. La synthèse de la parole permet de conserver un nombre réduit des paramètres influençant sur la parole.

II.2.2.Parole modifiée

Plusieurs techniques de modifications de la parole ont été testées :

1. Le filtrage passe-bas. Un filtre est appliqué de manière à ne laisser passer que les basses fréquences. Cependant, une partie des phonèmes reste reconnaissable, la technique suivante permet de réduire cette reconnaissance.
2. La resynthèse de parole. Des passages de paroles sont d'abord analysés pour extraire les phonèmes et leurs durées. Ces informations sont ensuite utilisées pour synthétiser de la parole, en éliminant certains informations, comme l'intonation ou certains phonèmes.
3. La réduction du spectre par vocodeur. Seulement une à quatre bandes spectrales sont disponibles.
4. Nous présentons les résultats obtenus avec ces techniques pour trois types de populations : adultes, nouveau-nés et primates non-humains.

II.2.2.1.Adultes

Le travail de Ramus et coll. (1999) s'appuie sur l'hypothèse de perception des voyelles par les adultes et également par les nourrissons, formulée par Mehler et coll., en 1996. Le signal de parole est représenté sous la forme d'une succession de consonnes et voyelles de durées variables. Afin de déterminer si les sujets peuvent discriminer les langues uniquement à partir du rythme de la parole, les stimuli présentés doivent contenir seulement des indices rythmiques. Plusieurs paramètres acoustiques, phonétiques et prosodiques sont mesurés pour décrire des énoncés. Les phrases analysées sont tirées du corpus LSCP [74] . Ces phrases sont segmentées en catégories phonétiques, en identifiant et en alignant chaque phonème par rapport au signal de parole, à l'aide d'un logiciel de visualisation du signal. Ensuite, une partie de ces informations est prise en compte pour synthétiser de nouveaux énoncés avec une voix artificielle. Cette resynthèse est effectuée à l'aide du logiciel MBROLA, qui utilise une base de diphones, dont la durée, la fréquence fondamentale et les phonèmes peuvent être spécifiés.

Ainsi, quatre transformations différentes sont évoquées. Pour chaque transformation, toutes les voyelles sont synthétisées par une unique voyelle [a] qui remplace toutes les voyelles :

1. Saltanaj : Dans cette transformation, les fricatives sont remplacées par [s], les liquides par [l], les occlusives par [t], les nasales par [n] et les semi-voyelles par [j]. L'intonation est copiée sur la courbe de F0 mesurée. Une partie de la phonotactique, le rythme et l'intonation sont préservés.
2. Sasasa : Toutes les consonnes sont remplacées par [s].
3. Sasasa plat : Même transformation que la précédente, avec un contour mélodique constant.
4. Aaaa : Cette transformation ne rend compte que du contour intonatif.

Ces transformations successives permettent d'isoler le rythme de la parole, et de « supprimer » l'intonation.

Les sujets doivent discriminer l'Anglais du Japonais, à partir des phrases resynthétisées. Les deux langues sont discriminées dans toutes les conditions exceptées pour celles ne contenant que le contour intonatif. Ainsi, les sujets ne distinguent pas les intonations anglaises et japonaises, si celles-ci sont dissociées de la composante rythmique créée par la succession des consonnes et des voyelles. Le rythme de la parole donné par la suite des consonnes et voyelles est suffisant pour discriminer l'Anglais du Japonais (condition **Sasasa plat**). En outre, les sujets échouent uniquement dans la condition **Aaaa,** seule condition où le rythme n'est pas présent. Des expériences précédentes (Maidment 1976, 1983 ; de Pijper, 1983) avaient montré que deux langues pouvaient être distinguées par leurs intonations, mais les locuteurs étaient natifs d'au moins une des deux langues à distinguer. Ramus (1999) propose donc une nouvelle

[74] 5 phases par locutrices ont été retenues, soit 20 phrases par langues, ce qui conduit à 160 phrases en tout. Les phrases sont appareillées en nombre de syllabes entre 15 et 19, et durent 3 secondes en moyenne.

expérience avec des sujets anglophones natifs. Ils obtiennent un score significativement supérieur au hasard, montrant que les intonations anglaises et japonaises peuvent être distinguées, dans la mesure où les sujets ont des connaissances préalables sur au moins une des deux langues. Que se passe-t-il pour les propriétés spectrales de la parole ?

Une seule étude a proposé d'étudier les propriétés spectrales lors de la discrimination des langues (Mori et coll., 1999). Lors de leur première expérience, les propriétés spectrales sont supprimées. Les sujets effectuent la distinction du Japonais et de l'Anglais avec un score de 85 % en se basant sur l'intonation et les propriétés rythmiques. Dans leur seconde expérience, le spectre est réduit entre une et quatre bandes spectrales. Ainsi, l'enveloppe temporelle est intacte, mais la F0 est supprimée. Les performances augmentent entre 1 et 4 bandes, de 63 % à 94 %. Seulement, il reste possible que certains mots soient identifiés clairement par les sujets, puisque la même expérience a été effectuée pour la reconnaissance des mots (Shannon et coll., 1995).

Les adultes sont sensibles à différentes propriétés de la parole, comment les nouveau-nés réagissent-ils à ces propriétés, alors qu'ils en ont très peu de connaissances ?

II.2.2.2.Nouveau-nés

Le filtrage passe-bas de la parole permet d'extraire la prosodie (Schaffer, 1984 ; den Os, 1988 ; Mehler et coll., 1988 ; Dehaene-Lambertz, 1995 ; Nazzi et coll., 1998). L'effet obtenu correspond à la parole traversant un mur. Ainsi certains phonèmes et syllabes, parfois des mots restent audibles. Ce filtre a été appliqué sur différentes langues (Nazzi et coll., 1998). Les nourrissons distinguent l'Anglais du Japonais, mais pas l'Anglais du Néerlandais, ils sont donc capables d'utiliser les propriétés prosodiques des langues pour les différencier, puisque la majorité des informations phonétiques et phonotactiques est rendue inexploitable par le filtrage.

Ainsi, la discrimination est possible uniquement quand les langues appartiennent à des classes rythmiques distinctes. En outre, ils ont prouvé qu'un changement de langue et de classe rythmique est plus significatif qu'un changement de voix. L'hypothèse de discrimination des langues en fonction de leur classe rythmique ne peut être confirmée pour les nouveau-nés dans la mesure où peu de langues ont été testées.

Les nourrissons discriminent également le Néerlandais du Japonais, sur la base du rythme produit par la succession des consonnes et des voyelles rendue par resynthèse vocale. Dans le cas de la parole naturelle, l'augmentation du nombre de locuteurs réduit les facultés de discrimination (Ramus, 1999). Jusczyk, Pisoni et Mullenix (1992) avaient postulé que la variabilité des voix induit un coût de traitement supplémentaire. Effectivement après une mémorisation de 2 minutes, les bébés de 2 mois ne pouvaient plus discriminer des syllabes, lorsqu'elles étaient prononcées par plusieurs locuteurs. L'utilisation de la resynthèse vocale permet d'éliminer les différences inter-locuteurs, tout en conservant les particularités rythmiques des langues étudiées, puisque les nouveau-nés distinguent le Néerlandais du Japonais lors de la resynthèse avec les classes SALTANAJ (Ramus, 2002b). Est-ce que les propriétés rythmiques de la parole peuvent être perçues par des primates non humains ?

II.2.2.3.Primates non-humains

Les singes tamarins ont également été testés sur des stimuli composés de parole resynthétisée. Les singes distinguent le Néerlandais du Japonais, avec la parole naturelle et les stimuli SALTANAJ. Cependant, Hauser (2002) suggère que les singes utilisent des indices acoustiques différents de ceux employés par les nouveau-nés.

L'amélioration d'un système d'IAL peut se réaliser par l'examen de nouvelles dimensions comme la grammaire ou la prosodie, sous exploitée actuellement (Zissman et Berkling, 2001). Des études perceptuelles ont tenté d'apporter des réponses quant au traitement de la prosodie, pour la discrimination ou l'identification des langues. Les indices acoustiques, et plus particulièrement la prosodie sont utilisés par les êtres humains (adultes et nourrissons) pour identifier une langue qui leur est inconnue. Comment peut-on traiter la prosodie pour qu'elle soit incorporée dans un système automatique destiné à l'IAL ?

II.3.Etat de l'art des études de la prosodie en IAL

Les systèmes d'IAL basés sur la prosodie se distinguent en deux catégories : ceux s'appuyant sur le rythme, et ceux mêlant la fréquence fondamentale, l'intensité et le rythme. En outre, un constat sur l'intégration de la prosodie dans les systèmes d'IAL sera proposé à la fin de cette section.

II.3.1.Le rythme seul

Les premières études dédiées au rythme de la parole emploient une segmentation manuelle du signal en consonnes et voyelles. Ces travaux ont naturellement été étendus par des systèmes de détection automatique des consonnes et des voyelles.

II.3.1.1.Segmentation manuelle en consonnes et voyelles

Ramus et coll. (1999) ont retenu 20 phrases pour chaque langue du corpus LSCP, en éliminant les phrases, dont la vitesse se différenciait trop des autres, afin d'éviter une normalisation du signal qui ne semble pas possible actuellement [75] (Ramus, 2002a). Trois variables ont été choisies pour caractériser le rythme :

1. %V, le pourcentage d'intervalles vocaliques, correspondant à la durée des intervalles vocaliquesToute séquence ininterrompue de voyelles., divisée par la durée totale de la phrase.
2. ∆V, l'écart-type des durées d'intervalles vocaliques au sein de la phrase.
3. ∆C, l'écart-type des durées d'intervalles consonantiquesToute séquence

[75] Une normalisation du débit de parole peut être obtenue lorsque la différence de durées entre deux segments (vocaliques ou intervocaliques) est divisée par la durée totale de ces deux segments (Grabe et Low, 2002). La moyenne de ces différences normalisées caractérise alors une phrase et permet de retrouver les grandes classes rythmiques de langues.

ininterrompue de consonnes. au sein de la phrase.

Les variables %V et ∆C permettent de regrouper les langues selon les trois grandes classes rythmiques classiques (syllabiques, accentuelles et moraïques).

Une régression logistique à partir de %V permet de discriminer des paires de langues. Les performances sont fonction des classes rythmiques, supérieures à 60 % dans le cas des langues de classes distinctes, inférieures à 60 % pour les langues d'une même classe. L'auteur ne propose pas d'identification des classes rythmiques elles-même. Les expériences de discrimination observées chez le nouveau-né sont simulées en appliquant un test de Mann-Whitney à la variable %V. Lorsque le test est inférieur à 0.05, les langues appartiennent à la même classe rythmique. Dans ce cas, l'Anglais et le Néerlandais sont bien distingués du Japonais, mais sont confondus entre eux (Ramus, 1999). Toutefois ces calculs ne permettent pas d'apporter de précision sur le mécanisme neurologique de traitement de la parole, qui effectue cette distinction.

Ces simulations faites à partir du pourcentage vocalique (Ramus, 1999) ont été également effectuées par Dominey et Ramus (2000) avec l'aide du réseau récurrent temporel TRN, décrit dans le chapitre 1. Les 20 phrases sont subdivisées en deux phases d'apprentissage et de validation de 10 phrases chacune, prononcées par quatre locuteurs distincts. Le réseau TRN reçoit en entrée la catégorie consonne ou voyelle par l'intermédiaire d'une des unités de la couche d'entrée. De plus, la durée n'est pas indiquée explicitement au réseau. Les classes rythmiques sont représentées soit par une seule langue (Anglais et Hollandais pour les langues accentuelles, et Japonais pour les langues moraïques), soit par une mixture de deux langues d'une même classe rythmique ou de classes rythmiques différentes. Les résultats obtenus concordent avec les études perceptuelles réalisées avec les nouveau-nés (Nazzi et coll., 1998). Ainsi, les langues d'une même classe rythmique ne peuvent être distinguées, contrairement aux langues de classes rythmiques distinctes. Un apprentissage non supervisé a également été testé. Les réseaux TRN entraînés avec l'Anglais, ont un temps de réaction plus long lorsqu'ils répondent pour les phrases Japonaises.

Tableau 3.2 Performances pour la discrimination des langues tirées de Dominey et Ramus (2000) et Ramus (1999).

Discrimination	Anglais / Japonais	Anglais / Hollandais
Classes rythmiques	Différente	Même
Nouveau-nés	P < .01	P = .16
Corrélation des états du TRN par rapport aux classes rythmiques	P < .001	P = .87
Performance du TRN	78 %	52 %
Régression %V	92.5 %	57.5 %
Adultes (condition sasasa plat)	68.1 %	-
Test de Mann-Whitney	P < .0001	P = 0.18

Les simulations précédentes proposent donc des méthodes pour tirer parti du rythme produit par les consonnes et les voyelles. Toutefois, il reste encore à déterminer comment

celles-ci peuvent être identifiées au sein du signal de parole.

II.3.1.2.Segmentation automatique

Les résultats obtenus par Mehler et coll. (1996) suggèrent que les nouveau-nés utilisent une perception approximative du signal fondée sur la sonorité. Galvès et coll. (2002) déterminent une fonction qui associe une partie du signal à une classe sonore (0 ou 1), de façon à segmenter la parole en deux grandes classes pseudo-phonétiques, proches des consonnes et des voyelles [76] . La valeur moyenne de la sonorité au cours d'une phrase joue alors le rôle du pourcentage moyen de voyelles. Les langues du corpus LSCP sont alors classées en fonction de leur classe rythmique sans faire appel à une segmentation manuelle. Leurs résultats sont semblables (excepté pour le Polonais) à ceux obtenus par Ramus et coll. (1999). Toutefois, leurs approches évitent de recourir à ces catégories phonétiques, aussi précises que les consonnes et les voyelles, auxquelles les nourrissons ne semblent pas avoir accès avant 6 et 9 mois.

L'Identification Automatique de 5 langues tirées du corpus MULTEXT a été réalisée à partir d'une caractérisation locale du rythme (Pellegrino, Chauchat, Rakotomalala et Farinas, 2002 ; Farinas, 2002). Un modèle du rythme de la parole doit tenir compte des informations segmentales (partie voisée/non-voisée de la parole), et suprasegmentales (organisation des unités rythmiques sur la totalité du discours). Les auteurs ont recours à la notion de pseudo-syllabe, pour obtenir une définition locale du rythme. Cette segmentation automatique sera employée pour le traitement global du rythme (décrite dans la section III.2.1). Elle sera utilisée pour les premières expérimentations (section IV.1 et IV.2).

La fréquence fondamentale est une autre information prosodique pertinente pour l'IAL, comment est-elle traitée dans les systèmes d'IAL ?

II.3.2.La fréquence fondamentale et l'intensité

Trois systèmes (deux statistiques et un connexioniste) ont fondé leur approche sur l'intonation et l'intensité pour identifier les langues.

Thymé-Gobbel et Hutchins (1996) ont étudié un certain nombre d'indices acoustiques [77] pour discriminer des paires de langues du corpus OGI-MLTS (Anglais, Espagnol, Japonais, Mandarin). La tonalité apparaît comme l'indice le plus efficace en particulier lorsqu'il est combiné avec la position de la phrase. D'après Laver (1994), une différence de 1 Hz dans la F0 peut être relevée, tandis que la limite se situe autour de 10 à 40 ms pour la durée des segments. Ainsi, la F0 est caractérisée par 430 niveaux perceptibles, alors que le rythme n'aurait que 25 catégories distinguables. Dans ce cas, l'intonation

[76] Il calcule l'entropie relative entre le spectre courant normalisé et les trois spectres normalisés précédents (fréquences entre 0 et 800 Hertz). La fonction de sonorité correspond à la moyenne de quatre entropies relatives consécutives.

[77] Le système mesure 224 motifs individuels : différence de F0 d'une syllabe à l'autre, par rapport au maximum, durée des syllabes, différence d'amplitude entre les syllabes, par rapport au maximum, transformé de Fourier pour les basses fréquences, nombre de syllabes par secondes, position de la phrase.

offre un plus grand nombre de variations possibles entre les langues. Les performances oscillent entre 73 % et 83 %, mais sont maximales pour une combinaison différente d'indices fonction des langues à discriminer.

Itahashi et Du (1995) proposent une analyse discriminante de 21 paramètres concernant F0 et l'intensité (écart-type, skewness et kurtosis, corrélation ente la F0 et intensité, etc.). 6 langues du corpus OGI-MLTS ont été retenues, et les locuteurs sont tous des hommes. Seul le score d'apprentissage est indiqué : 63.3 % d'identification pour 20s de signal. Itahashi, Kiuschi et Yamamoto (1999) présentent les performances de leur système fondé sur la F0 pour 10 langues, celles-ci atteignent 37 % en apprentissage et 28 % en validation, des scores supérieurs au hasard.

Cummins et coll. (1999) utilisent des réseaux récurrents pour réaliser la discrimination de 10 paires de langues du corpus OGI-MLTS à partir du décours temporel de F0 et de l'intensité. Pour l'enveloppe de l'intensité, les résultats varient entre 72.2 % et 50.7 %. Pour la trajectoire de F0, les performance s'échelonnent de 47.4 % à 73.2 %. Le Français se détache clairement de toutes les autres langues. Le Japonais et le Mandarin sont extrêmement confondus. Lorsque les deux indices (F0 et intensité) sont combinés les performances sont globalement [78] améliorées et varient de 48.8 % à 78.2 %. Les taux de reconnaissance sont ensuite représentés sous forme d'un arbre, qui rassemble entre elles les langues dont les scores sont proches. Ainsi le Mandarin, l'Espagnol et le Français apparaissent ensemble, comme représentant des langues syllabiques, et l'Anglais et l'Allemand, comme caractéristiques des langues accentuelles.

Considérés seuls, les systèmes d'IAL fondés sur la prosodie n'approche pas les performances des systèmes phonotactiques. Quel peut être l'apport d'un système prosodique à un système d'IAL existant ?

II.3.3.Intégration de la prosodie dans un système d'IAL

La combinaison de la prosodie avec des systèmes d'IAL porte ses fruits principalement pour les durées inférieures à 10 secondes. Hazen et Zue (1997) proposent un modèle prosodique s'appuyant sur les parties voisées de la F0, et sur la durée des segments utilisés dans leur modèle phonotactique. Itahashi et coll. (1999) ont combiné les paramètres de F0 à des coefficients sceptraux. Les performances progressent de 96.7 % à 97.3 %. Thymé-Gobbel et Hutchins (1996) ont testé avec succès leur modèle d'IAL avec trois paires de langues contenant l'Anglais (Zissman et Martin, 1995). Cependant, cette intégration pourrait être améliorée en modifiant l'architecture classique d'un système d'IAL, de façon à laisser une place prépondérante à la prosodie (Leavers et Burley, 2001).

Effectivement, pour répondre au problème d'identification des langues, les êtres humains intègrent diverses composantes. Dans le cas d'une langue connue, ils utilisent des stratégies impliquant leurs connaissances de la langue : stratégie lexicale [79] comme

[78] Pour le Mandarin et le Japonais il apparaît que le réseau ne sait pas « trier » les entrées pour utiliser seulement l'amplitude qui permet une meilleure distinction de ces langues, que lorsque la F0 et l'intensité sont combinées.

[79] Cette dernière stratégie est la plus rapide et la plus sûre des méthodes présentées, mais aussi la plus complexe à réaliser.

l'identification de mots individuels. Face à une langue étrangère, ils font appel à des connaissances non-linguistiques : stratégie suprasegmentale (le rythme, les accents et les contours intonatifs) et stratégie segmentale (les caractéristiques phonétiques ; Leavers et Burley, 2001). Ces deux dernières stratégies peuvent servir de base à une nouvelle architecture pour l'IAL. Contrairement à la conception classique, l'identification des langues se ferait par regroupement, une classe de langues serait privilégiée avant d'identifier spécifiquement la langue utilisée.

Un exemple précis permet d'expliciter cette stratégie (Leavers et Burley, 2001). Quatre langues (Chinois, Anglais, Espagnol et Portugais) doivent être identifiées (Figure 3.3). L'aspect suprasegmental de l'intonation permet d'isoler le Chinois. Dans le cas des langues avec des tons lexicaux, la fréquence fondamentale n'est influencée que localement par la coarticulation des tons voisins. En revanche, pour les langues ne possédant pas de ton lexical, la tonalité est influencée par la phrase entière. Cette caractérisation globale est obtenue par le coefficient de Hurst [80] . Des propriétés segmentales (les voyelles) distinguent l'Espagnol du Portugais ce qui influe sur la forme [81] de la distribution de F0 ainsi que sur la durée moyenne des segments voisés.

Cet article démontre que la cognition et la linguistique peuvent s'insérer dans la réalisation d'un système d'IAL, afin d'utiliser des paramètres pour isoler différentes classes de langues. Toutefois, ce test est effectué sur seulement 4 langues. Le même travail de regroupement pour 11 langues (voire 120 langues) s'avérerait beaucoup plus fastidieux. Effectivement, il faudra sans doute tenir compte de langues ayant une classe à part comme le Basque, ou bien appartenant à plusieurs classes différentes.

[80] Ce coeffcient, noté H, est égal à 0.5 pour une série aléatoire. Pour les autres valeurs, il s'agit d'une chaîne d'éléments interdépendants. Pour le Chinois 0<=H< 0.5 (balancement entre les tons lexicaux, haut puis bas), et 0.5<H<=1 dans les autres cas (continuité, bas puis haut). En outre, ce paramètre est indépendant des locuteurs.

[81] Skewness : coefficient d'asymétrie.

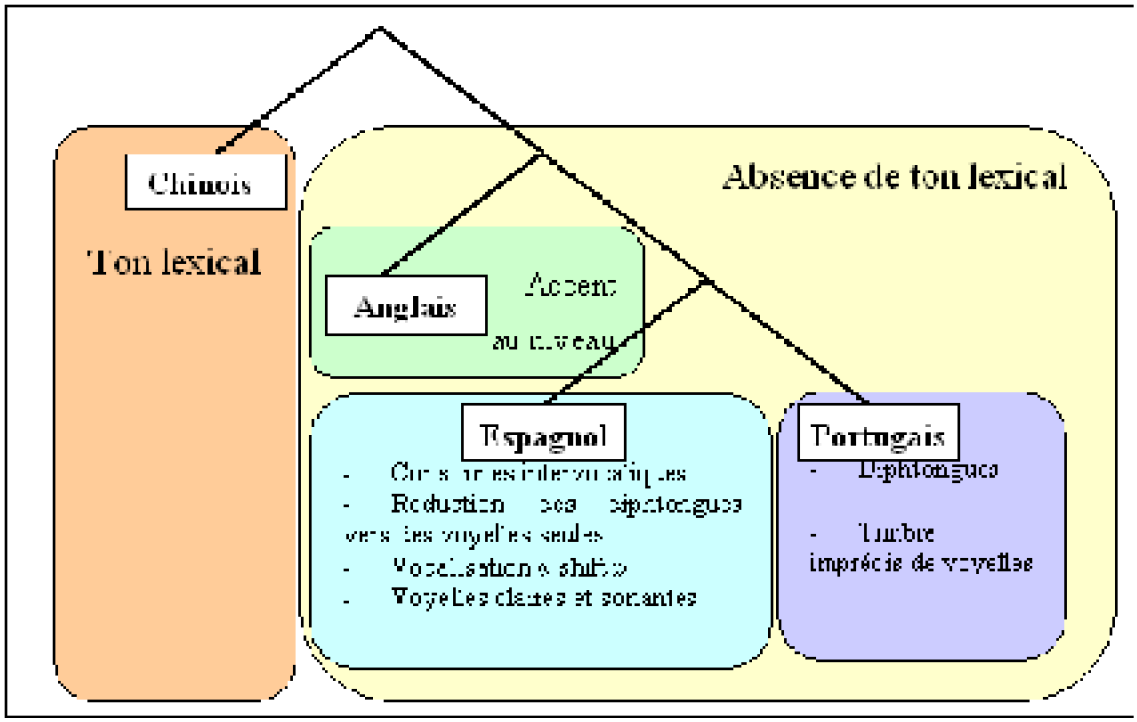


Figure 3.3 Illustration de l'identification de quatre langues d'après Leavers et Burley (2001).

Notre but est donc d'aborder l'identification automatique des langues à partir des composantes acoustiques, sans faire appel à des connaissances linguistiques. Chaque passage à identifier sera donc retranscrit comme des séquences temporelles. Ces séquences temporelles seront traitées par le réseau récurrent temporel décrit précédemment (chapitre 1 : section III.3). La section suivante résume les méthodes et les corpora employés dans les expérimentations.

III.Matériel et méthodes

Cette section décrit dans un premier temps les corpora contenant les langues à identifier, puis un descriptif des méthodes employées pour décrire le signal acoustique.

III.1.Corpora

Trois corpora ont été étudiés dans cette section. Les deux premiers ont été utilisés dans une perspective d'identification automatique des langues. Le dernier corpus LSCP a été utilisé pour rendre compte des propriétés du traitement temporel effectué par le réseau TRN. Effectivement, nous emploierons ce corpus pour tester les capacités du réseau TRN en fonction des classes rythmiques des langues à discriminer.

III.1.1.Le corpus MULTEXT

Les expériences conduites en IAL ont été testées essentiellement sur le corpus

MULTEXT. Une des tâches du projet MULTEXT (Llisterri, 1996) consistait à fournir un étiquetage prosodique de la base de données de parole multilingue EUROM1 (projet ESPRIT SAM 2589). Les passages composés de phrases thématiquement liées entre elles, issus du 'Few Speaker Set' d' EUROM1, constituent le matériel de base de notre étude, car ils proposent une cohérence linguistique et prosodique. Pour chaque langue, 10 locuteurs se répartissent la lecture de 40 passages (5 phrases par passage), ce qui correspond à une centaine de passages par langue (500 phrases).

Ce corpus contient **5 langues européennes** (Anglais, Français, Allemand, Italien, Espagnol). Ce corpus a servi pour l'IAL dans l'étude de Pellegrino et coll. (2002) et la thèse de J. Farinas (2003). Les enregistrements ont été obtenus dans des conditions de laboratoire et ne contiennent donc pas de bruits parasites, tels que des rires ou des hésitations. Ce corpus contient dix locuteurs par langues, ce qui représente au final 749 passages différents. Les caractéristiques sont apprises à partir de la moitié du corpus MULTEXT (374 séquences de 5 secondes pour 5 langues, cf. Tableau 3.4). Les performances indiquées proviennent de l'autre moitié du corpus, composée de locuteurs, qui ne sont pas présents dans le corpus d'apprentissage.

Langues	Nombre de locuteur	Nombre de passages par locuteur	Durée totale (min.)	Durée d'un passage (s.)
Anglais	10	15	44	17,6
Français	10	10	33	21,9
Allemand	10	20	77	21,9
Italien	10	15	54	21,7
Espagnol	10	15	52	20,9
TOTAL	50	75	4h19	20,7

Tableau 3.3 Caractéristiques du corpus MULTEXT d'après Pellegrino et coll. (2002).

	Anglais	Français	Allemand	Italien	Espagnol	TOTAL
Apprentissage	75	50	100	75	75	375
Validation	75	49	100	75	75	374
TOTAL	150	99	200	150	150	749

Tableau 3.4 Nombre de passages par langue pour le corpus MULTEXT, divisé en corpus d'apprentissage et de validation

III.1.2.Le corpus OGI-MLTS [82]

Il s'agit du premier corpus spécifiquement dédié à l'IAL. Ce corpus est composé uniquement de discours non contraint et est utilisé comme référence dans les publications ayant trait à l'IAL. Les travaux les plus récents en IAL s'appuient généralement sur ce corpus pour pouvoir comparer leurs résultats entre eux. Il a été considéré comme base lors de la première évaluation des systèmes d'IAL promulguée par NIST (mars 1993) à laquelle ont participé 8 sites de recherches américains. De cette base constituée de 22 langues, nous n'avons retenu que 6 langues (Anglais, Japonais, Espagnol, Mandarin, Hindi, et Allemand). Il s'agit des langues extraites de l'extension (OGI-MLTS 22 langues)

[82] **Multi Language Telephone Speech.**

du corpus initial.

Chacune de ces langues est décrite par la durée de chaque phonème ainsi que par la catégorie du phonème rencontrée. Cette catégorisation est accomplie manuellement par des experts linguistiques. Ces groupes prennent en compte un certain nombre de bruit n'appartenant pas au langage à proprement parlé. De cette description phonémique, il n'est retenu que les trois catégories consonnes, voyelles et silences. Eventuellement une classe supplémentaire sera considérée pour tous les bruits non verbaux. Chaque langue comporte un ensemble de locuteurs (*cf.* Tableau 3.5), qui expriment une courte histoire (notée *story_bt.wav)* dans leur propre langue par l'intermédiaire du téléphone. Ce texte peut varier de quelques secondes jusqu'à 45 secondes.

	Anglais	Japonais	Allemand	Espagnol	Hindi	Mandarin	TOTAL
Apprentissage	68	43	47	48	38	40	275
Validation	67	43	54	60	38	27	282
TOTAL	146	64	101	108	68	70	557

Tableau 3.5 Nombre de passages par langue pour le corpus OGI-MLTS

III.1.3.Le corpus LSCP

Il s'agit d'un corpus multilingue (18 langues en tout), créé à l'origine par Thierry Nazzi (Nazzi et coll., 1998). Ce corpus comporte 54 phrases écrites en Français, et traduites approximativement dans chacune des autres langues. Ces phrases sont déclaratives et typiques de bulletins d'informations radiophoniques. Elles ont été lues par 4 locutrices natives de chaque langue qui énoncent chacune 5 phrases. Chacune de ces phrases est relativement courte : autour de 2 à 3 secondes de signal de parole. Huit d'entre elles ont été segmentées en consonnes et voyelles par F. Ramus. Nous avons étudié les paires de langues utilisées dans les expérimentations chez le nourrisson (Nazzi et coll., 1998), pour simuler ces expériences en employant le réseau TRN pour traiter directement une représentation spectrographique des basses fréquences du signal acoustique (section IV.4).

III.2.Représentation des données

Les données transmises au TRN sont des représentations du signal acoustique, obtenues par une détection automatique des consonnes et des voyelles, un cochléogramme, ou des représentation spectrographique des basses fréquences.

III.2.1.Le rythme

Dans le cas du corpus MULTEXT, le rythme est donné par la succession des consonnes et des voyelles, détectées automatiquement par l'algorithme développé par Pellegrino (1998). Les unités rythmiques de base sont obtenues à partir de l'algorithme « Forward-Backward Divergence » [83] . Les pauses sont mises de côté grâce à un

[83] Les segments les plus courts correspondent aux explosions (burst) et aux parties transitoires, et les plus longs aux zones stables des voyelles.

détecteur d'activité vocale. Les voyelles sont reconnues à partir de l'énergie contenue dans les fréquences basses du signal, indépendamment du locuteur et de la langue [84]. Chaque passage d'une langue est traduit par une liste contenant la succession des consonnes, des voyelles et des silences éventuels, ainsi que la durée de chacun de ces événements. Le corpus OGI-MLTS a été fourni avec une segmentation en consonnes et voyelles obtenue manuellement.

La couche d'entrée du réseau TRN est composée de deux neurones, l'un codant les consonnes, l'autre les voyelles. Si une voyelle dure 80 ms, le neurone codant la voyelle sera activé pendant 16 itérations. Les silences sont traduits par une absence de signal d'entrée.

III.2.2.Cochléogramme

Le cochléogramme fournit un modèle de la cochlée, qui est l'outil de décomposition spectrale de l'oreille. Les passages de parole sont codés par une collection de trames de 256 composants chaque 10 ms (une matrice unique calculée par le logiciel PRAAT [85] contenant l'intonation (F0) et l'information spectrale décrite en fonction du temps). La résolution fréquentielle est fixée à 10 Barks, la taille de la fenêtre d'analyse est de 30 ms, la fenêtre de masquage rétrograde est de 30 ms également. Ces valeurs sont les valeurs par défaut du logiciel PRAAT.

III.2.3.Représentation spectrographique des basses fréquences

Trois représentations (calculées à partir du logiciel PRAAT) ont été testées pour transmettre les fréquences les plus basses au réseau TRN. Pour des raisons techniques [86], le cochléogramme ne sera pas utilisé. Les deux premières mesures donnent un vecteur de valeurs toutes les 5 ms, et utilisent une fenêtre d'analyse large pour obtenir les valeurs des basses fréquences, voisines de la fréquence fondamentale. En outre, ces deux représentations intègrent des informations sur l'intensité du signal pour chaque bande de fréquence, entre 0 et 400Hz, que ne fournit pas la dernière méthode :

1. Valeurs des filtres alignés sur une échelle de perception Mel (algorithme Melfilter de PRAAT) inférieure à 500 mels avec une fenêtre d'analyse de 60 ms. La résolution est de 12.5 mels ce qui conduit à 40 neurones d'entrées (Figure 3.13).
2. Un spectrogramme avec une fenêtre d'analyse de 80 ms en tenant compte que des valeurs inférieures à 400Hz. La résolution fréquentielle est fixée à 2.75 Hz, et conduit à 145 neurones d'entrées. Chaque valeur est multipliée par 500 (Figure 3.14 gauche).
3. Un calcul de la valeur de F0 seule à partir d'une méthode d'autocorrélation. Cette

[84] Seuls les silences de plus de 150 ms sont pris en compte. Tout ce qui n'est pas reconnu comme une voyelle est alors assimilée à une consonne.

[85] Logiciel dédié à l'analyse du signal de parole (disponible gratuitement sur internet : http://www.praat.org)

[86] Le logiciel PRAAT ne permet pas d'obtenir uniquement les basses fréquences avec le cochléogramme.

méthode fournit initialement une valeur numérique de F0. Cette valeur est alors représentée à l'aide d'une population de 60 neurones dont l'activité est calculée à partir d'une courbe de Gauss dont la forme est fixe (Figure 3.14 droite, cf. chapitre 4 section III.3).

Dans tous les cas, les valeurs de chaque bande de fréquences sont comprises entre 0 et 100. Si une valeur d'activation est supérieure à 100, elle est fixée à 100, avant d'être transmise au réseau. La plage des valeurs transmises au réseau est donc toujours comprise entre 0 et 100. En outre, ces deux représentations intègrent des informations sur l'intensité du signal pour chaque bande de fréquence, entre 0 et 400Hz, que ne fournit pas la dernière méthode.

III.3.Méthodes de traitement

Pour étudier l'Identification Automatique de Langues, nous avons eu recours à deux types de méthodes : statistiques et connexionistes.

III.3.1.Méthodes statistiques

L'objectif des méthodes suivantes est de proposer une premier traitement du rythme par des moyens statistiques. Dans un premier temps, nous proposerons d'étudier le comportement du pourcentage vocalique au cours du temps. A un temps *t* donné, celui-ci correspond à la proportion d'intervalles vocaliques (toute séquence ininterrompue de voyelles ; Ramus, 1999) au sein d'un segment temporelle [0, *t*]. Il s'agit de la durée totale des intervalles vocaliques, contenus dans le segment [0, *t*] divisée par le temps *t*. Ensuite, un classifieur à moyenne gaussienne sera aussi employé, pour identifier les langues à partir de la durée des consonnes et des voyelles, et pour obtenir une base des performances, par une méthode de classification, faisant référence en statistique. La variable étudiée sera le pourcentage vocalique. Un passage de parole sera représenté par une séquence contenant *n* pourcentages vocaliques, correspondant aux *n* premiers phonèmes. Un descriptif plus approfondi de cette technique sera présenté dans la seule section expérimentale où elle est employée (IV.1.2).

III.3.2.Méthodes connexionnistes

Pour compléter ces études, nous proposons d'employer le réseau TRN décrit dans le chapitre Un (section III.3). Dans ce chapitre, trois méthodes ont été testées avec le TRN.

La première méthode est appliquée à chaque pas de la simulation aux états des couches State et $State_D$ du réseau TRN. Elle est dénommé « Une Trame ». Effectivement, l'analyse ne prend en compte que le vecteur issu de ces états à un instant *t*. Ces vecteurs sont ensuite utilisés comme un encodage de la séquence traitée par le TRN. L'apprentissage s'effectue à partir d'un prototype moyen de ces vecteurs (cf. chapitre Un section III.3.1). Les évènements passés sont donc encodés dans la couche de contexte $State_D$ uniquement.

La seconde méthode Accumulation emploie la totalité des vecteurs formés entre *0* et

l'instant courant *t*. A chaque pas de la simulation, un compteur est incrémenté pour chaque langue par la distance entre le vecteur obtenu à l'instant *t* pour le passage à identifier et les prototypes de chaque langue à cet instant *t* (*cf.* Figure 3.4). Pour éviter que les valeurs des compteurs ne deviennent trop importantes, leur valeur minimale est retranchée à chacun d'eux. Le compteur ayant la plus petite valeur indique alors quelle est la langue reconnue. Cette méthode revient à tenir compte de matrices de taille *t*, au lieu de vecteurs pour décrire le signal. Un prototype correspond donc à la moyenne des matrices de chaque passage de la base d'apprentissage d'une langue. Cette méthode sera dénommée Accumulation et sera comparée à la méthode précédente (dénommée Une Trame).

La troisième méthode est une validation croisée employée pour augmenter le nombre de données disponibles en apprentissage. Un seul locuteur est testé lors de la validation, les autres forment la base d'apprentissage. Cette technique est très coûteuse pour effectuer une sélection du meilleur réseau lors de l'apprentissage. Elle sera alors réalisée avec le réseau TRN, qui a donné les meilleures performances en validation, lorsque l'apprentissage est réalisé avec la première moitié du corpus MULTEXT.

Ce matériel et ces méthodes ont constitué la base des expériences suivantes, réalisées dans le cadre de l'IAL.

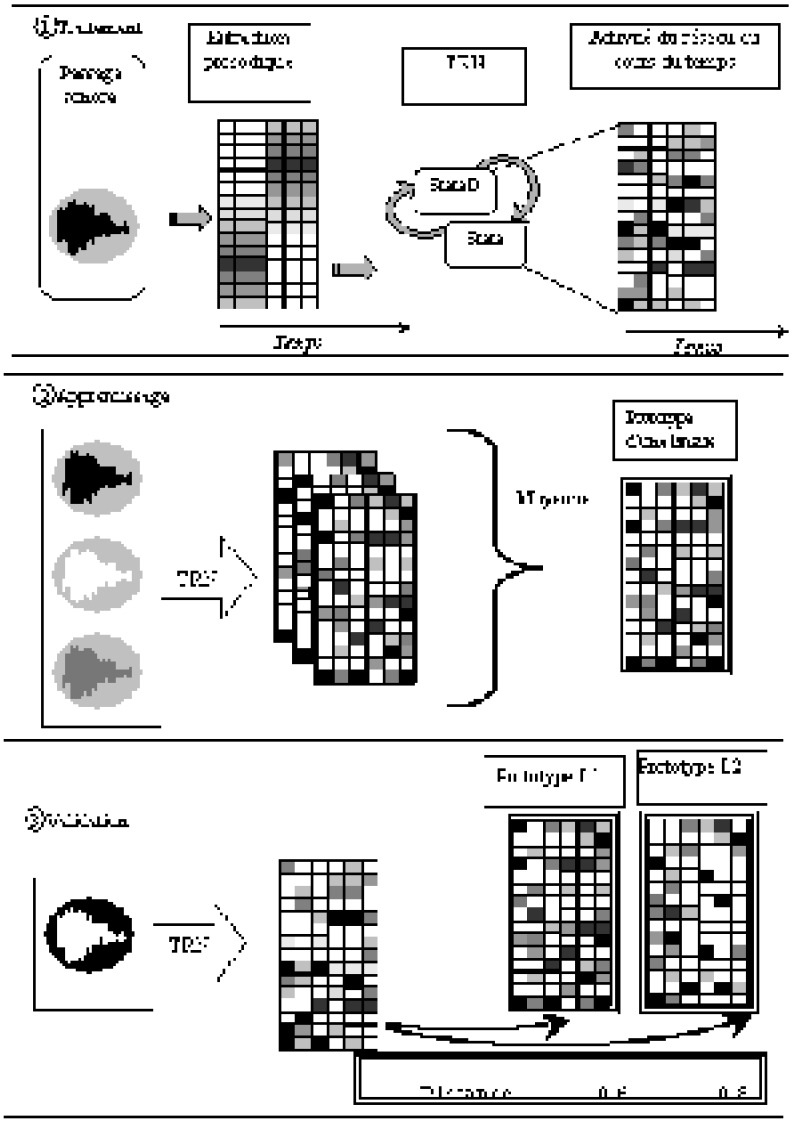


Figure 3.4 Le réseau TRN et la méthode d'accumulation.

IV.Expérimentation

Ce travail expérimental a pour objectif de montrer que le réseau TRN peut traiter globalement plusieurs dimensions présentes dans le signal acoustique de parole, pour contribuer à résoudre le problème de l'IAL. D'abord notre étude s'appuiera sur le rythme généré par la succession des consonnes et des voyelles, enfin nous transmettrons une représentation spectrographique du signal au réseau TRN. En IAL, notre travail se déroulera en quatre phases :

1. Les deux premières sections seront consacrées à l'IAL à partir du rythme engendré par la succession des consonnes et des voyelles. Les langues sont donc identifiées

uniquement à partir du rythme :

2. La première section (section IV.1) traitera d'une analyse statistique de la proportion d'intervalle vocalique au cours du temps.

3. A partir de ces observations nous appliquerons le réseau TRN (section IV.2) aux séquences formées par les consonnes et les voyelles.

4. La section suivante (IV.3) abordera une représentation plus complète de la prosodie, sans segment ni distinction entre les dimensions de la F0, des formants et du rythme.

5. Enfin, nous proposons de modéliser l'expérience de Nazzi et coll. (1998), pour la discrimination des langues, en fonction de leurs classes rythmiques (section IV.4). Nous tiendrons compte de la représentation employée précédemment pour les fréquences inférieures à 400 Hz. Effectivement, les données de l'expériences avec les nourrissons étaient soumises à un filtre passe-bas.

6. L'objectif de ce chapitre est donc double : 1. nous devons montrer que le réseau peut contribuer à l'IAL en respectant une contrainte temporelle, 2. cette même contrainte temporelle doit pouvoir influer sur la discrimination des langues, en fonction de leur classe rythmique, de façon à refléter des propriétés de traitement observées chez le nourrisson.

IV.1.Approche Statistique du Rythme en IAL

Le but de cette section est d'étudier l'évolution au cours du temps du pourcentage d'intervalles vocaliques, défini en section III.3.1. Nous appliquerons un classifieur à moyenne gaussienne, pour identifier les langues, à partir d'une caractérisation globale du rythme.

IV.1.1.Pourcentage d'intervalles vocaliques au cours du temps

Ramus et coll. (1999) ont montré que le rapport de la durée de l'ensemble des voyelles par la durée de la totalité des consonnes pour une phrase est caractéristique d'une langue. Ce paramètre permettait de retrouver les trois grandes classes rythmiques des langues (accentuelles, syllabiques et moraïques). Le graphique suivant (*cf.* Figure 3.5) prend en compte l'évolution au cours du temps de ce paramètre pour chacune des 5 langues européennes du corpus MULTEXT. Il s'agit de la moyenne de ce rapport à travers l'ensemble des phrases disponibles de ce corpus. L'axe des ordonnées indique la valeur du pourcentage vocalique pour une durée inférieure ou égale à l'abscisse.

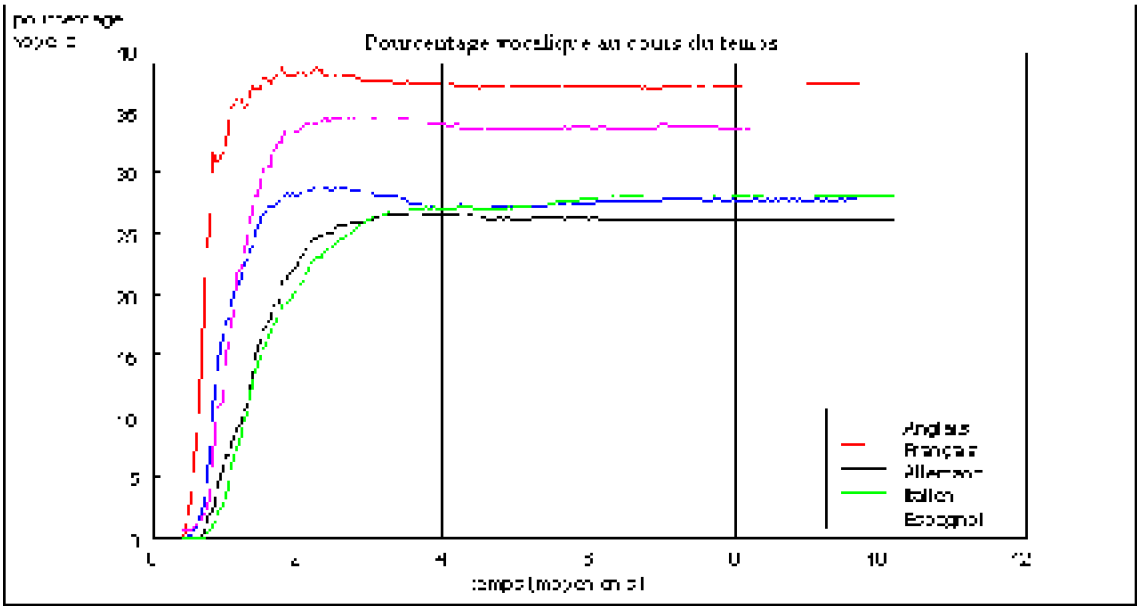


Figure 3.5 Evolution du pourcentage vocalique au cours du temps pour le corpus MULTEXT.

Ce graphique (*cf.* Figure 3.5) montre que les langues ne présentent pas la même progression au cours du temps. Par exemple, entre 1 et 2 secondes l'Anglais se démarque de l'Italien alors qu'à partir de 4s de signal les deux langues semblent se confondre. En tenant compte de l'évolution temporelle du pourcentage vocalique il serait plus facile de distinguer certaines langues à un moment précis dans le temps. Est-ce que cette évolution au cours du temps peut être utilisée pour l'identification automatique des langues ?

Les valeurs du pourcentage vocalique se stabilisent autour de 4s pour les corpora OGI-MLTS et MULTEXT. A travers ces corpora, les classes rythmiques des langues sont respectées. Les langues accentuelles (Allemand, Anglais, Hindi) ont un pourcentage vocalique moins élevé que les langues syllabiques (Espagnol, Français), qui ont un pourcentage vocalique inférieur aux langues moraïques comme le Japonais. Le Chinois Mandarin n'a pas de classe rythmique reconnue. L'Italien a un comportement plus proche de la classe accentuelle (Figure 3.5).

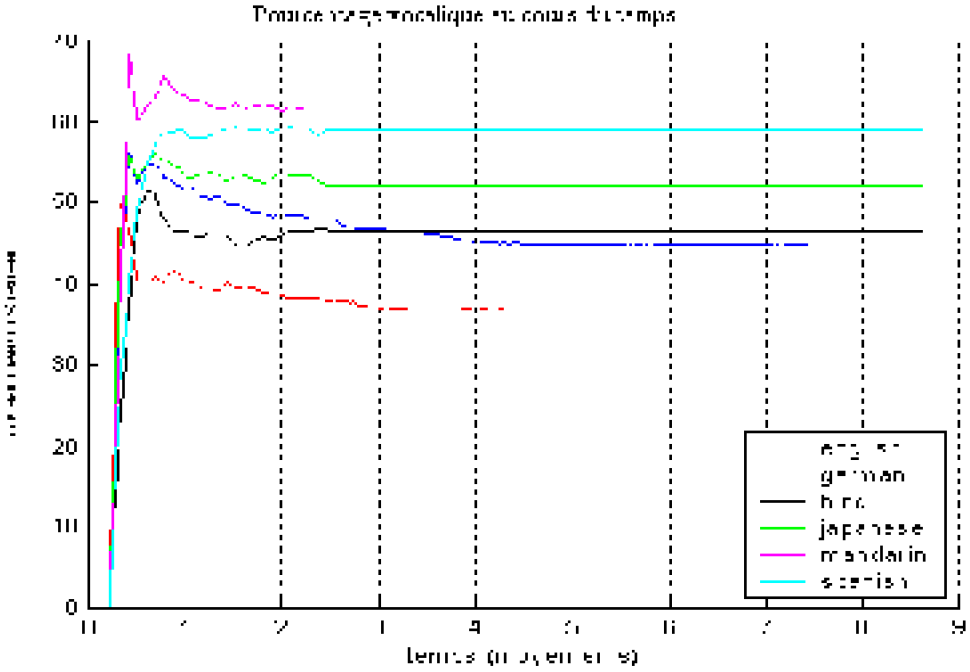


Figure 3.6 Evolution du pourcentage vocalique au cours du temps pour le corpus OGI-MLTS.

Ces graphiques (Figure 3.5 et 3.6) indiquent que le pourcentage de segment vocalique subit une évolution propre à chaque langue. Il doit donc être possible de prendre en compte cette évolution, pour identifier automatiquement les langues.

IV.1.2.Classifieur à moyenne gaussienne

Nous décrivons ici une méthode statistique d'identification automatique des langues à partir du pourcentage vocalique.

IV.1.2.1.Descriptif de la méthode employée

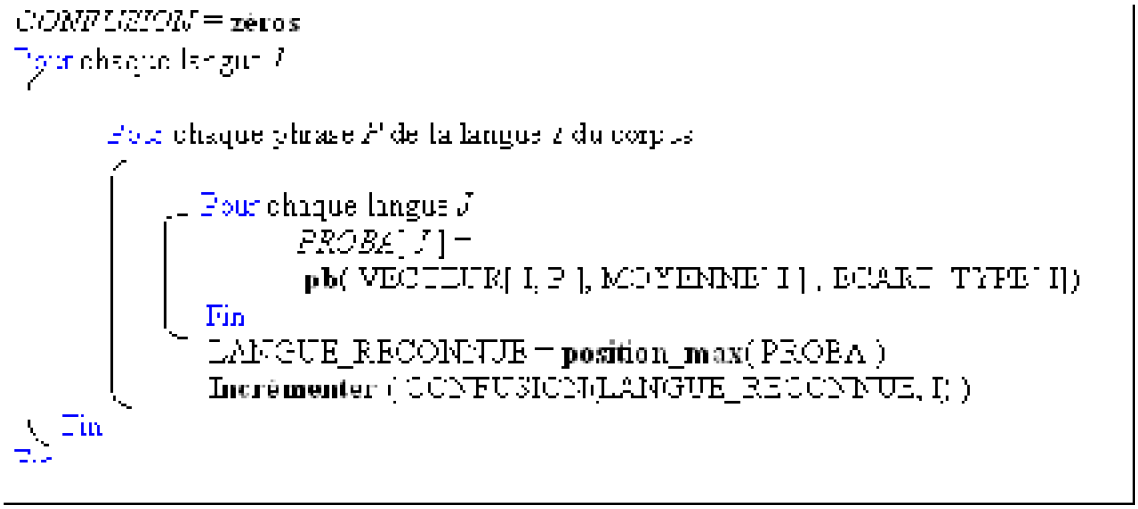
L'algorithme prend en compte un nombre fixe de phonèmes, qui doit être le même pour chaque passage à identifier. Une phrase est représentée par un vecteur indiquant le pourcentage vocalique obtenu sur la durée du début de la phrase jusqu'au phonème courant.

Les tests sont effectués avec un corpus d'apprentissage et un corpus de validation, tirés du corpus OGI-MLTS. Pour éviter une dissymétrie des corpora, ils seront testés chacun pour la phase d'apprentissage et de validation. La moyenne des résultats obtenus sur les corpora utilisés en validation ou en apprentissage est alors indiquée.

IV.1.2.2.Algorithme

La première étape consiste à établir un modèle gaussien (moyenne et matrice de covariance) pour chaque langue en accumulant les caractéristiques rythmiques traduites par le pourcentage vocalique au cours du temps. Il en ressort ainsi un gabarit pour

chaque langue testée. Ces valeurs étant obtenues, chaque séquence exprimée est comparée avec ces deux références, par l'utilisation de la courbe de Gauss, pour chacune des langues. Au bout du compte, la probabilité la plus élevée est celle de la langue reconnue.



Algorithme 3.1 Classification par moyenne gaussienne.

La fonction **position_max** renvoielaposition de la valeur maximale seulement dans le cas où celle-ci est unique. La fonction **zéros** initialise une matrice à 0. Ensuite, les vecteurs *MOYENNE* et *ECART_TYPE* contiennent respectivement la moyenne et l'écart-type des valeurs prises pour la durée d'énonciation des phonèmes. La matrice *CONFUSION* indique comment les langues sont reconnues, et quelles langues prêtent à confusion. Enfin, la fonction **pb** calcule la valeur suivante de la probabilité gaussienne :

$$pb(X,M,C) = \frac{e^{\frac{1}{2}(x-m)C^{-1}(x-m)^T}}{\sqrt{(2\pi)^n \det(C)}}$$ Equation 1

Avec :

- *X* le vecteur induit par une phrase correspondant aux valeurs du pourcentage vocalique ;
- *C* la matrice de covariance de l'ensemble des valeurs pour une langue donnée ;
- *M* le vecteur moyenne de l'ensemble des valeurs pour une langue donnée ;
- *n* la dimensiondu vecteur *M.*

IV.1.2.3.Résultats

Nb Phonèmes	30	60	90
Durée totale (s)	3	6	9
Apprentissage	50%	63%	73%
Validation	20%	25%	26%

Tableau 3.6 Performance en fonction du nombre de phonèmes qui entrent dans le calcul de la moyenne (6 langues du corpus OGI-MLTS).

L'augmentation du nombre de phonèmes permet d'améliorer les performances pour l'apprentissage et la validation. Les performances en validation dépassent le seuil du hasard. Ces premières expériences montrent qu'il existe des différences rythmiques entre les langues, mais qu'il est difficile d'obtenir un modèle statistique global satisfaisant pour l'identification des langues. Est-il possible d'appliquer un modèle issu des neurosciences et sensible aux propriétés temporelles pour identifier les langues à partir du rythme, engendré par les consonnes et les voyelles ?

IV.2.Identification des Langues par le Rythme avec le Réseau TRN

Dorénavant, nous utiliserons le réseau TRN pour identifier les langues (décrit dans le chapitre 1 section III.3). La sélection du réseau s'effectue à partir d'une population de 50 individus, dont seuls les poids des connexions diffèrent. Dans un premier temps, nous appliquerons le réseau TRN à la totalité du signal de parole, codé en consonnes et voyelles. Ensuite, nous examinerons les performances obtenues au cours du temps, avant de proposer une méthode supplémentaire au réseau pour améliorer les performances. Enfin, nous appliquerons une méthode de validation croisée.

IV.2.1.Premier résultat

Le réseau TRN reçoit en entrée le type de phonèmes (consonnes ou voyelles) par l'intermédiaire de deux neurones. Seuls l'ordre sériel et la durée des consonnes et voyelles peuvent être exploités par le réseau TRN. Le réseau encode sous la forme de vecteurs les passages de parole, codés en consonnes et voyelles. Cette première évaluation permet de comparer les performances d'identification du modèle avec le traitement statistique mis en place par l'IRIT (Farinas et André-Obrecht, 2000). Le corpus OGI-MLTS a en effet été testé avec la même liste de phrase pour l'apprentissage et la validation.

Les résultats prennent en compte la totalité des langues disponibles pour chaque corpus, et n'exploitent que le rythme induit par la succession des consonnes et des voyelles. Pour le corpus MULTEXT de validation, les performances sont de l'ordre de 30 %. Pour le corpus OGI-MLTS, le taux d'identification est de 33 %, alors que les méthodes statistiques précédentes atteignaient seulement 26%. En outre, Farinas et Obrecht (2000) obtenaient un taux de l'ordre de 34 % avec une approche statistique locale. Ainsi le modèle neuromimétique se comporte de la même manière avec le corpus OGI-MLTS qu'avec un traitement statistique. Cependant, certains paramètres du modèle

n'ont pas été optimisés pour une tâche d'identification des langues, ce qui laisse présager que les performances peuvent être améliorées.

IV.2.2.Evolution des performances au cours du temps

Le graphique (Figure 3.7) présente l'évolution des résultats au cours du temps, obtenus pour le corpus OGI-MLTS. L'apprentissage est effectué à partir des vecteurs obtenus pour une durée de 5s. Toutes les 5 secondes, l'état courant du réseau est comparé aux relevés de l'état du réseau obtenus à 5s, pendant l'apprentissage. Les performances sont maximales lorsque la durée des fichiers tests correspond à la durée des fichiers appris (5s) (pic à 40 % dans la figure 3.7). De même, lorsque l'apprentissage est effectué avec les données correspondant à 20s, les performances sont maximales à 20s. Les performances restent stables autour de ce maximum, et au dessus du hasard (30 % vs. 17 %).

L'architecture des réseaux récurrents privilégie l'information des évènements récents, une partie de l'information est donc perdue au cours du temps. Le paragraphe suivant présente une technique additionnelle au réseau mise en œuvre pour apporter une mémoire auxiliaire efficace au réseau récurrent TRN.

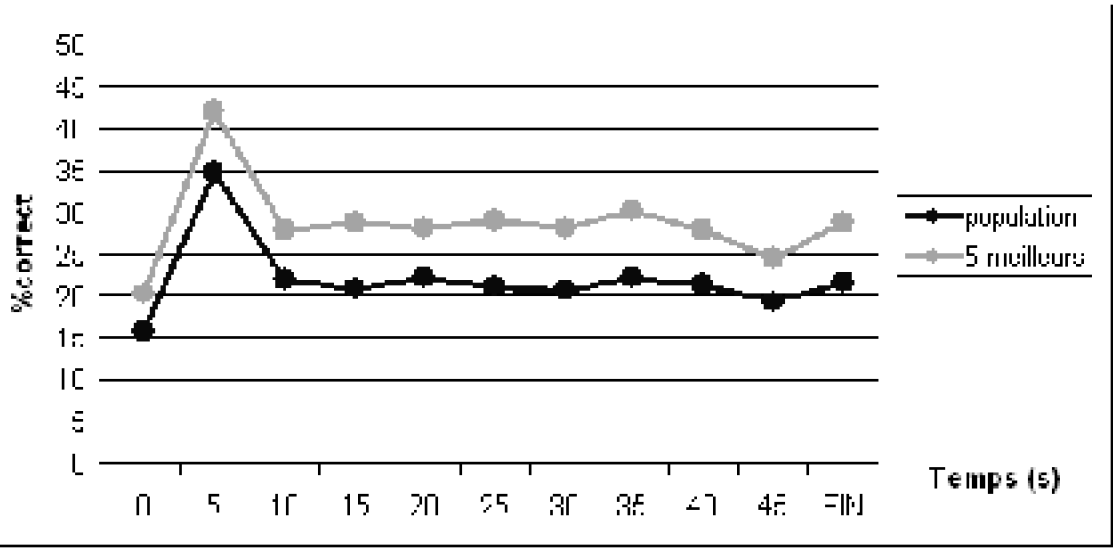


Figure 3.7 Apprentissage effectué pour 5s. Population indique la moyenne des scores de 50 réseaux, 5 meilleurs désigne le score moyen des 5 meilleurs réseaux en apprentissage. L'apprentissage est effectué uniquement avec les vecteurs extraits au bout d'une durée de 5s des passages (corpus OGI-MLTS).

IV.2.3.Accumulation des états d'activation du réseau

La méthode utilisée pour évaluer le réseau est décrite en section III.2.3 (*cf.* également Figure 3.4). Un prototype correspond à la moyenne des matrices de chaque passage de la base d'apprentissage d'une langue. Cette méthode permet de fournir une mémoire supplémentaire au réseau TRN pour encoder les informations passées.

L'utilisation d'une mémoire auxiliaire pour l'accumulation des informations provenant

du TRN augmente les performances de 33 % à 50.4 % pour un durée inférieure à 8 secondes (Figure 3.8). Cependant, les performances pour des durées supérieures à 8s sont décroissantes. Cela peut être lié à la diminution du nombre d'échantillon pour ces durées. Le réseau sélectionné est celui donnant les meilleures performances en validation pour une population de 50 réseaux.

IV.2.4.Méthode de validation croisée

Nous avons appliqué une méthode de validation croisée (leaving one out) au réseau TRN, pour augmenter la base d'apprentissage (cf. III.2.4). La validation croisée augmente les performances, et permet d'approcher un peu plus les performances obtenues avec les méthodes statistiques (Pellegrino et coll., 2002). Nous n'avons pas testé cette méthode dans le cadre d'une véritable validation aveugle (i.e. en effectuant la sélection pour chaque corpus d'apprentissage), car cette technique est trop gourmande en temps de calcul. Cette expérience prouve que l'ajout de connaissance dans la base d'apprentissage permet d'améliorer les performances.

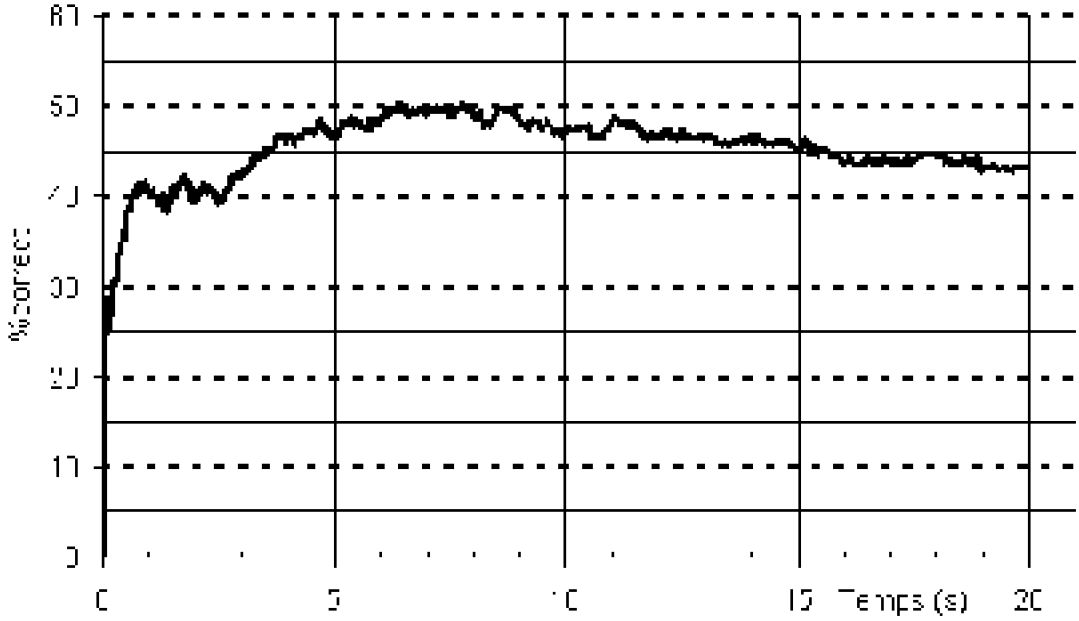


Figure 3.8 Performance du réseau TRN avec la méthode d'accumulation, présentant les meilleures performances en validation, sur le corpus MULTEXT.

Nous avons montré que les langues du corpus MULTEXT peuvent être identifiées à partir d'une segmentation automatique en consonnes et voyelles, en employant un réseau récurrent qui respecte une contrainte temporelle. Seulement, cette information est bien loin de refléter la complexité du signal acoustique. En particulier, l'intonation, l'intensité et le timbre n'apparaissent pas dans cette représentation. Est-il possible d'intégrer ces dimensions pour l'IAL ?

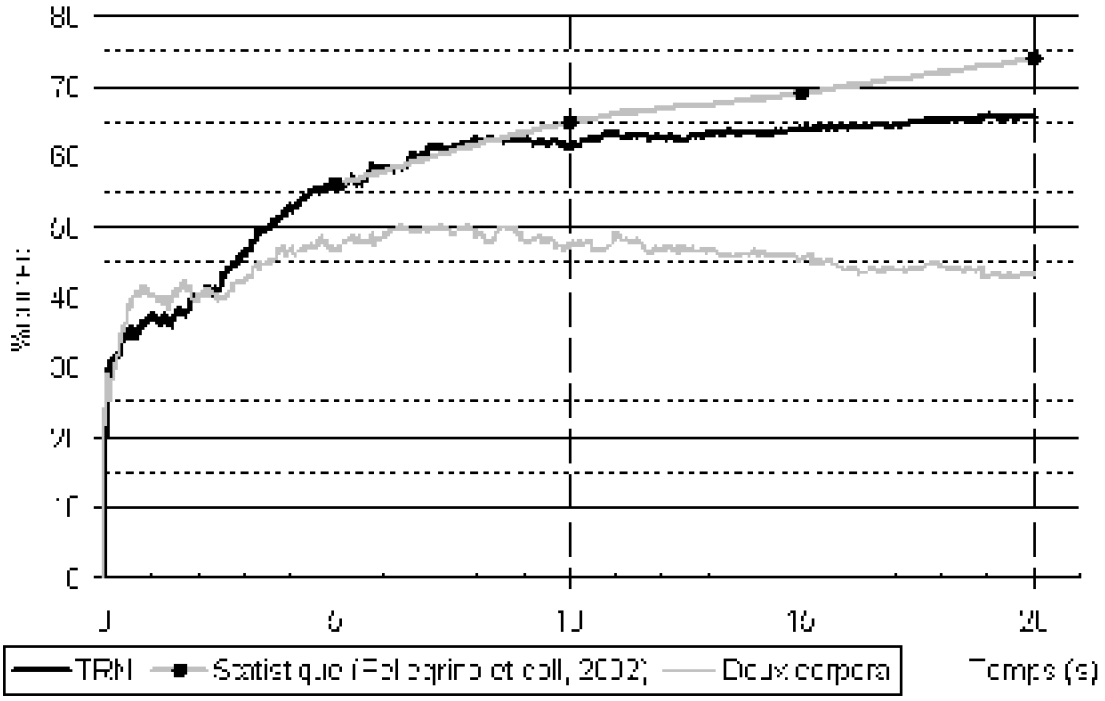


Figure 3.9 Performance du réseau TRN avec une méthode de validation croisée (TRN), les performances des méthodes statistiques avec la pseudo-syllabe (Pellegrino et coll., 2002), et les résultats du TRN avec l'apprentissage basé sur une division en deux du corpus MULTEXT (cf. section IV.2.3).

IV.3.Représentation acoustique non segmentée

Afin de compléter les travaux précédents, nous allons étudier une représentation spectrographique du signal acoustique. Cette représentation comporte :

1. une description temporelle très fine (une trame toutes les 10ms, qui représente le signal compris dans une fenêtre de 30ms) ;
2. des dimensions prosodiques (F0, F1, F2, et F3) qui ne sont pas isolées ;
3. Nous étudierons d'abord les résultats obtenus à partir d'un prototype moyen, avant de combiner le réseau TRN avec cette nouvelle représentation.

IV.3.1.Cochléogramme

Nous proposons donc maintenant d'utiliser une autre représentation du signal de parole, à savoir une description de la cochlée (ou cochléogramme, voir Figure 3.10), qui est l'outil de décomposition spectrale de l'oreille. L'excitation de la membrane basilaire dans l'oreille interne est modélisée au cours du temps, en réponse à un échantillon de parole. Le cochléogramme accentue les basses fréquences qui sont consacrées à la prosodie, contrairement à un spectrogramme classique. Ainsi la parole est codée par une collection de trames de 256 composants chaque 10 ms. Un passage de langue est entièrement

décrit par une matrice unique calculée par le logiciel de PRAAT (http://www.praat.org) contenant l'intonation (F0) et l'information spectrale décrite en fonction du temps.

Nous envisageons deux expériences :

1. Les performances d'identification des cinq langues du corpus MULTEXT ;
2. Les performances en discrimination de trois paires de langues du corpus OGI-MLTS.

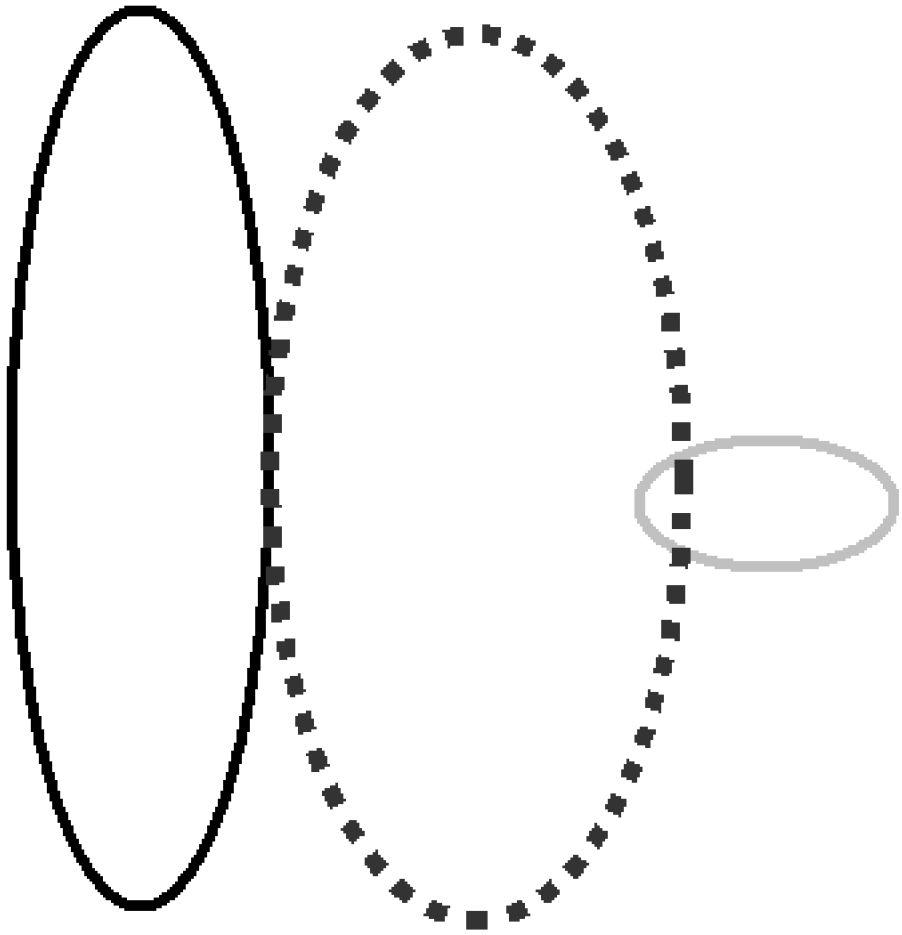


Figure 3.10 Cochléogramme d'un extrait de parole (en abscisse, le temps par pas de 10ms), en ordonnée les fréquences, 256 unités d'entrées du réseau) Le cercle en trait continu indique le bruit de fond. Le cercle pointillé indique la parole. Le cercle grisé indique une respiration.

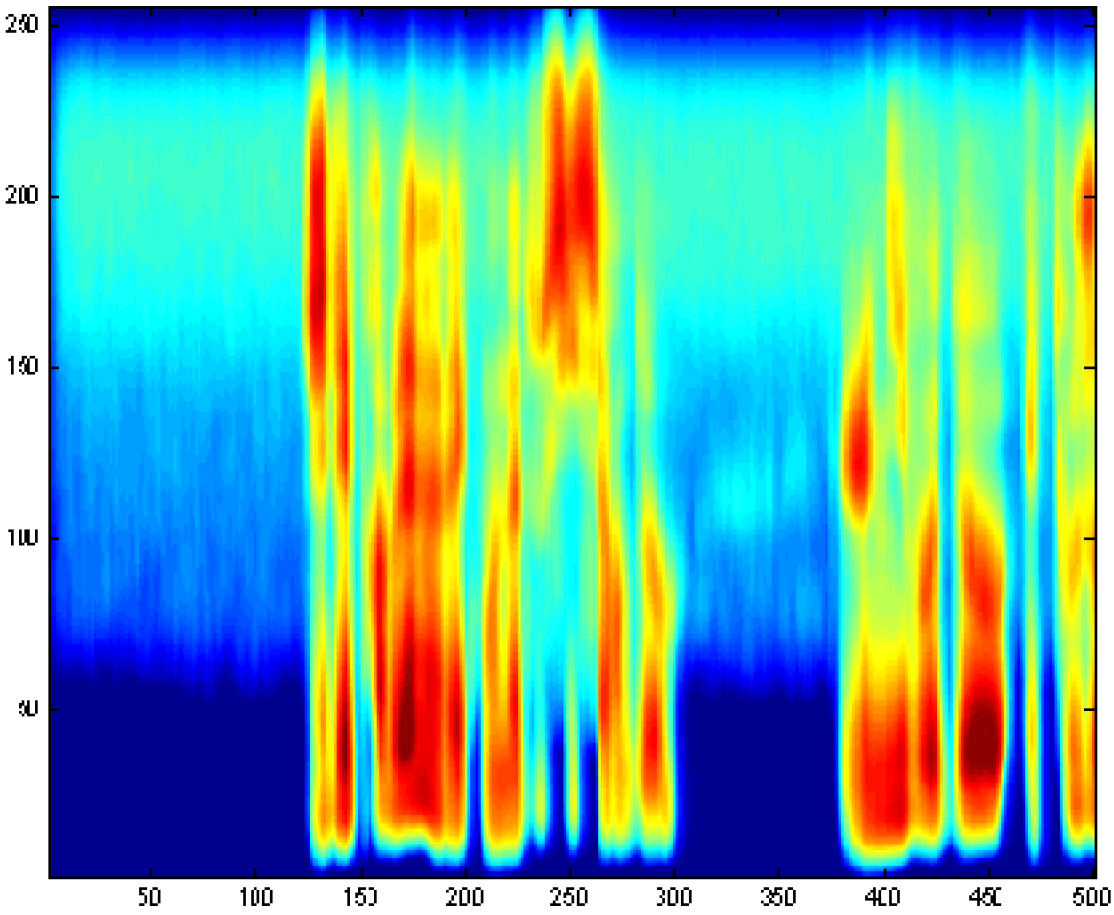


Figure 3.10 Cochléogramme d'un extrait de parole (en abscisse, le temps par pas de 10ms), en ordonnée les fréquences, 256 unités d'entrées du réseau) Le cercle en trait continu indique le bruit de fond. Le cercle pointillé indique la parole. Le cercle grisé indique une respiration.

Afin de tester la validité de cette représentation, la même expérience que précédemment a été réalisée sur le corpus MULTEXT à partir des fichiers entiers contenant la totalité du signal acoustique. Cette première expérience (Figure 3.11) montre que le bruit de fond est suffisamment corrélé avec les langues à identifier, pour obtenir un taux d'identification de 60 % alors que le signal de parole n'est pas encore présent. Les performances sont illustrées pour les deux méthodes Accumulation et Une Trame (i.e. une seule trame de signal est considérée) mais sans le TRN.

Dans un second temps, nous avons testé le cochléogramme sur trois paires de langues du corpus OGI-MLTS. Nous retrouvons de bonnes performances dans le cas de la distinction Allemand / Anglais (96 %) et également pour l'Hindi et l'Anglais (81 %) et ce pour la première trame du signal. En revanche, la distinction entre l'Hindi et l'Allemand ne présente pas de bonnes performances ni pour la première trame (42 %), ni lorsque 5 secondes de signal sont prises en compte (55 %).

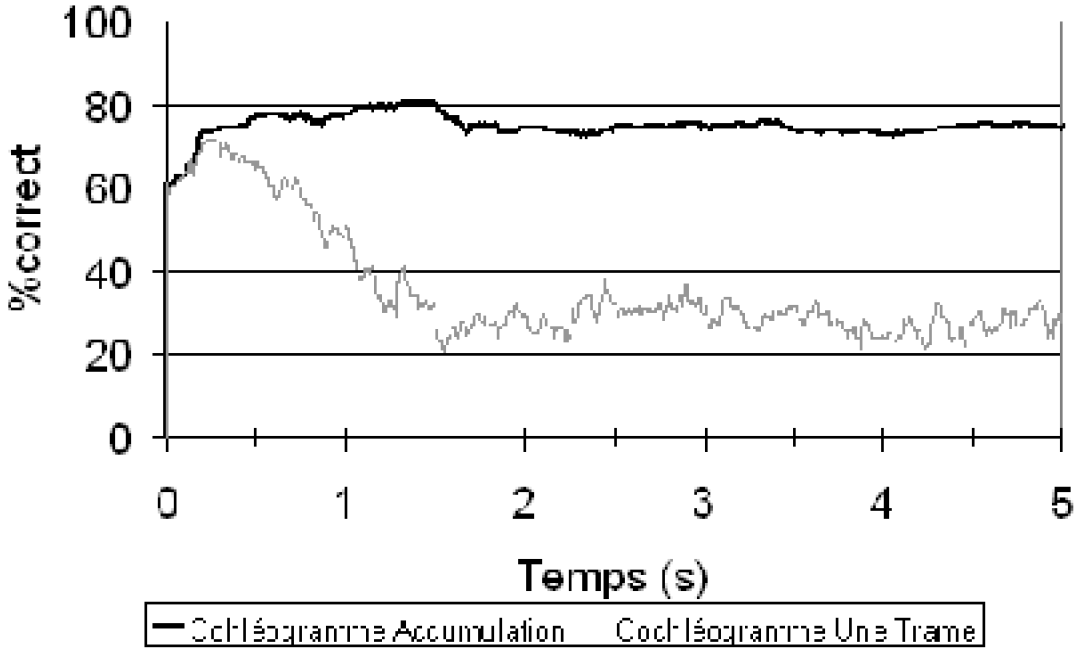


Figure 3.11 Performance pour deux méthodes de représentation du spectre.

IV.3.2.Utilisation conjointe du réseau (TRN) et du cochléogramme

Quel peut être l'impact du TRN sur le cochléogramme ?

Les deux premières secondes du signal de parole vont être éliminées afin de supprimer les indices (comme le bruit de fond), qui ne pourront être utiles dans une application réelle.

Notre objectif est maintenant d'examiner la capacité du TRN pour traiter les entrées calculées par le cochléogramme. Cette expérience a deux points à démontrer : 1) le TRN encode le signal spatio-temporel de parole dans une représentation spatiale de sorte qu'elle permette une identification comparable à une méthode sans fondement biologique (Accumulation); 2) cette représentation a pu être réinjectée dans cette même méthode pour augmenter les performances.

Nous devons évaluer la distance d'un échantillon donné d'une langue à chacun des prototypes, générés pour chaque langue apprise. Deux méthodes (appelées Une Trame et Accumulation) ont été examinées pour dissocier l'information comprise dans une seule trame, de celle contenue dans l'ensemble des trames décrivant le signal de parole.

Dans la première méthode (Une Trame), nous considérerons seulement une unique trame, pour représenter le signal de parole à un instant *t*. Ainsi le contexte ne peut pas être codé, lorsque le TRN n'est pas employé.

Dans une deuxième méthode, la distance de chaque trame individuelle sera additionnée pour tenir compte de tous les événements précédents entre *0* et l'instant *t* (Accumulation, présentée dans le paragraphe IV.2.3). Pendant la phase d'apprentissage un prototype sera établi, soit à partir d'une seule trame dans le cas de la méthode Une

Trame, soit à partir d'une matrice (Accumulation). Les résultats sont ceux obtenus pendant la phase de validation pour le réseau TRN, sélectionné dans une population de 50 réseaux lors de l'apprentissage.

Les résultats sont présentés dans la Figure 3.12. Une analyse de variance des taux d'identification de chaque langue a été conduite entre les trois facteurs : Durée (1s ; 5s ; 10s), Méthode (Simple ; Accumulation), codage (avec ou sans TRN). Il n'y avait aucun effet principal significatif, seule l'interaction entre la durée et les deux autres facteurs (méthode et codage) est significative. Ceci suggère que les deux méthodes d'évaluation et le codage s'exécutent différemment pour au moins une durée.

Un premier test post hoc LSD conduit sur la durée et la méthode a montré que la méthode Accumulation diffère de manière significative de la méthode Une Trame ($P < 0,01$) pour toutes les durées. Un deuxième test post hoc LSD conclut que les facteurs Durée et Codage montrent une différence significative entre l'absence et la présence du TRN pour 5s et 10s ($P < 0,01$), et non pour 1s ($P = 0,32$). Les deux analyses précédentes ont prouvé qu'il n'y avait aucune différence entre 5s et 10s (resp. $P = 0,69$; $P = 0.73$). En outre il n'y avait aucune différence significative entre le TRN et la méthode d'Accumulation appliquée aux entrées du TRN pour 5s ($P = 0,71$) ou de 10s ($P = 0,68$).

Ainsi les événements passés sont encodés par le TRN. Pour 10s, le mélange des deux méthodes temporelles (Accumulation et TRN) diffère du traitement avec uniquement le TRN ($P = 0,02$) ou uniquement la méthode Accumulation ($P = 0,04$), mais pas pour une durée de 5s ($P = 0,11$; $P = 0.06$). Ceci suggère que l'adjonction d'une mémoire, apportée par la méthode Accumulation est efficace pour des durées supérieures à 5s. Il n'y avait aucune différence significative entre 5s et de 10s pour le réseau le plus efficace, cependant les résultats ont progressé de 62,8 % à 64,9 %.

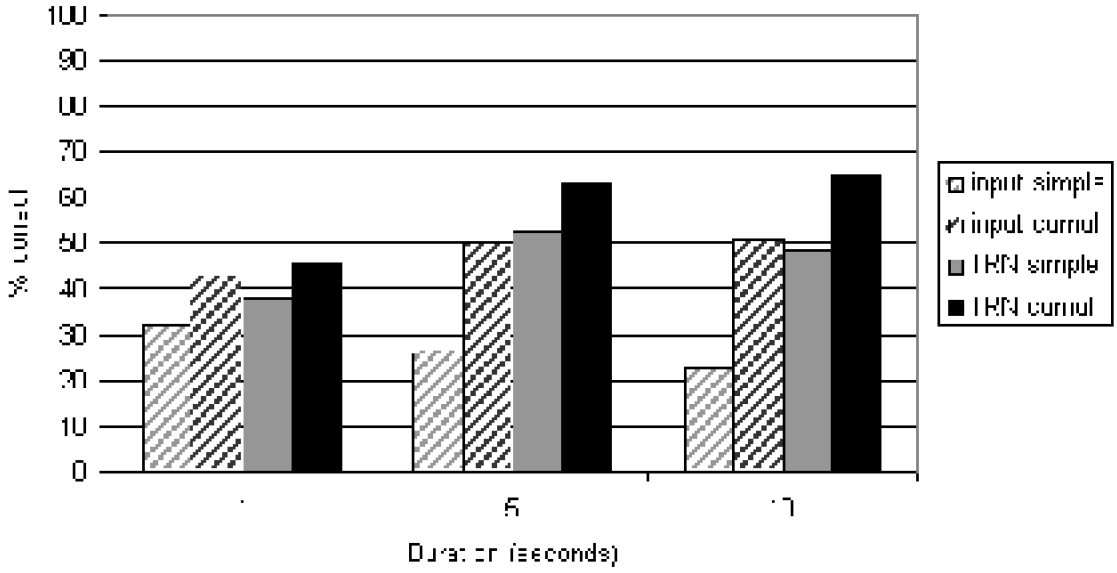


Figure 3.12 Performances d'identification des cinq langues européennes du corpus MULTEXT. Quatre méthodes sont présentées : avec ou sans le réseau récurrent (resp. TRN ou input), ainsi qu'avec ou sans la méthode d'accumulation (resp. simple ou cumul).

IV.4.Simulation de discrimination de langues

Les expériences précédentes ont permis d'adapter le réseau TRN au problème de l'IAL. Notre objectif est maintenant de tester si le réseau TRN permet de retrouver les résultats donnés par les expériences perceptuelles lors de la discrimination de langues de classes rythmiques différentes (Nazzi et coll., 1998). Effectivement, il a été démontré que le réseau TRN est sensible à cette différence rythmique, lorsque les langues sont représentées par une succession de consonnes et voyelles (Dominey et Ramus, 2000). Cette différence liée à la contrainte temporelle est-elle toujours perçue par le réseau TRN avec une représentation spectrographique des basses fréquences ?

Dans cette section, trois langues extraites du corpus LSCP (Nazzi et coll., 1998) seront utilisées : Japonais (langue moraïque), Anglais, et Néerlandais (langues accentuelles). Seul l'encodage donné par le réseau TRN à la fin d'une phrase sera retenu. La méthode d'accumulation ne sera pas employée. Les performances sont indiquées pour une population de 10 réseaux. En outre, l'information transmise au réseau sera contenue dans les basses fréquences, comme pour l'expérience réalisée avec les nourrissons (Nazzi et coll., 1998).

Trois représentations (calculées à partir du logiciel PRAAT [87]) ont été testées pour transmettre la prosodie des phrases au réseau TRN :

1. Filtres alignés sur une échelle de perception Mel (algorithme Melfilter de PRAAT, Figure 3.13)
2. Un spectrogramme à bande étroite avec une fenêtre d'analyse de 80 ms en tenant compte que des valeurs inférieures à 400Hz (Figure 3.14 gauche).
3. Un calcul de la valeur de F0 seule à partir d'une méthode d'autocorrélation. (Figure 3.14 droite).

Dans tous les cas, toutes les valeurs d'activations supérieures à 100 sont ramenées à 100. Les deux premières mesures donnent un vecteur de valeurs toutes les 5 ms, et utilisent une fenêtre d'analyse large pour obtenir les valeurs de fréquences basses, voisine de la fréquence fondamentale.

[87] Logiciel dédié à l'analyse du signal de parole (disponible gratuitement sur internet : http://www.praat.org)

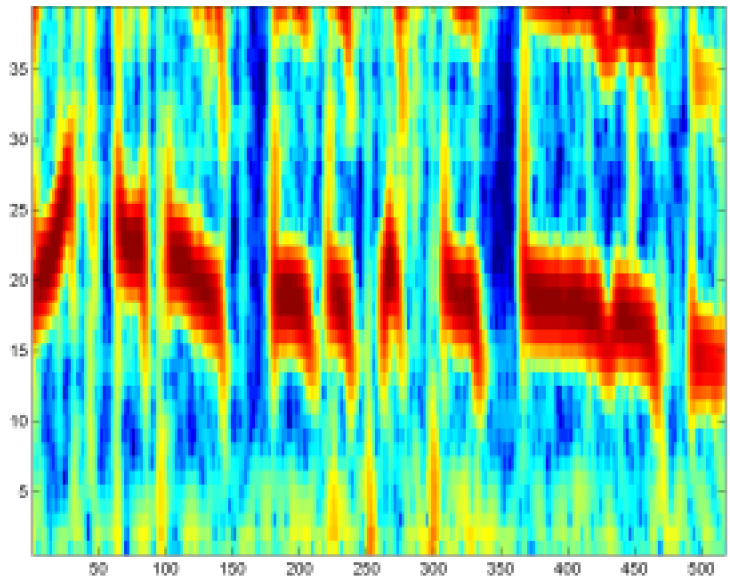


Figure 3.13 Représentation de F0 par l'algorithme Melfilter.

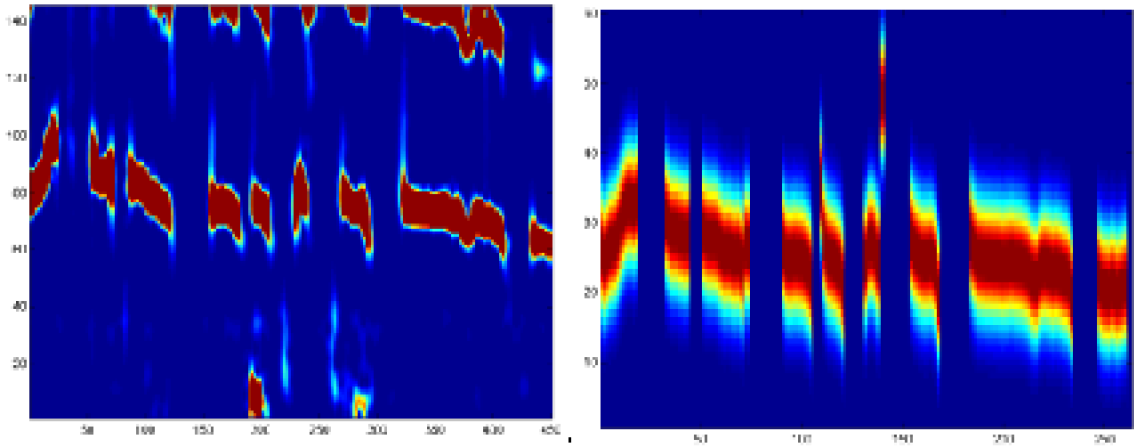


Figure 3.14 Représentation de F0, à gauche par un spectrogramme, à droite à partir d'une méthode d'autocorrélation, donnant une valeur numérique de F0, qui est représentée à l'aide d'une courbe de Gauss fixe.

La figure 3.15 présente la moyenne des performances obtenues par 10 réseaux récurrents sur le corpus de validation pour identifier les paires de langues Anglais/Japonais et Anglais/Néerlandais. Quelle que soit la méthode utilisée, la discrimination entre l'Anglais et le Japonais est plus aisée qu'entre l'Anglais et Néerlandais. Pour les représentations Melfilter et le spectrogramme, les performances dépendent des classes rythmiques des langues, comme pour les nourrissons (Nazzi et coll., 1998) (ANOVA [88] : 1) Melfilter : p=0.001, F=14.7 ; 2) Spectrogramme p<0.001, F=22).

[88] ANOVA avec un seul facteur Classe rythmique (identique ou différente).

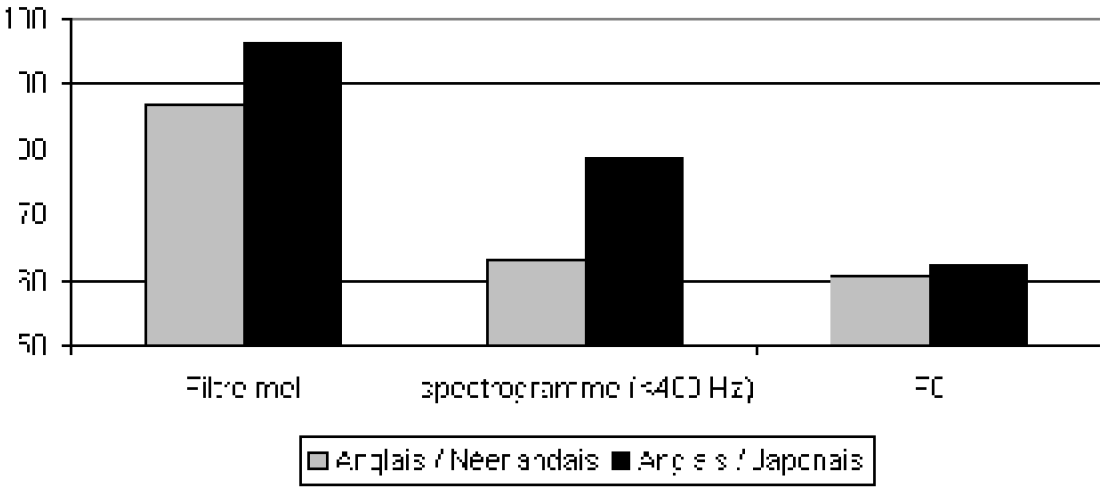


Figure 3.15 Performances de discrimination des langues en fonction des différentes descriptions de la prosodie.

Lorsque la F0 est transmise seule, les réseaux ne parviennent pas à distinguer les langues quelle que soit leur classe rythmique (ANOVA : 3) p=0.77, F=0.09). Les performances sont faibles dans le cas où seule la F0 est décrite au cours du temps (62 % et 63 %, le seuil du hasard est à 59 %). Lorsque l'intensité est disponible (spectrogramme à bande large), les performances augmentent (78 %), mais uniquement dans le cas ou les langues appartiennent à des classes rythmiques distinctes.

La première représentation utilisée donne de bonnes performances quelles que soient les conditions de classe rythmique. De surcroît, les performances sont supérieures à 80 % dès la première trame de signal. Elle inclut donc des informations supplémentaires. Cette représentation étant fortement bruitée, un seuil variable est appliqué de manière à ne tenir compte que des valeurs qui dépassent ce seuil. Plus le seuil est élevé, plus le signal est appauvri.

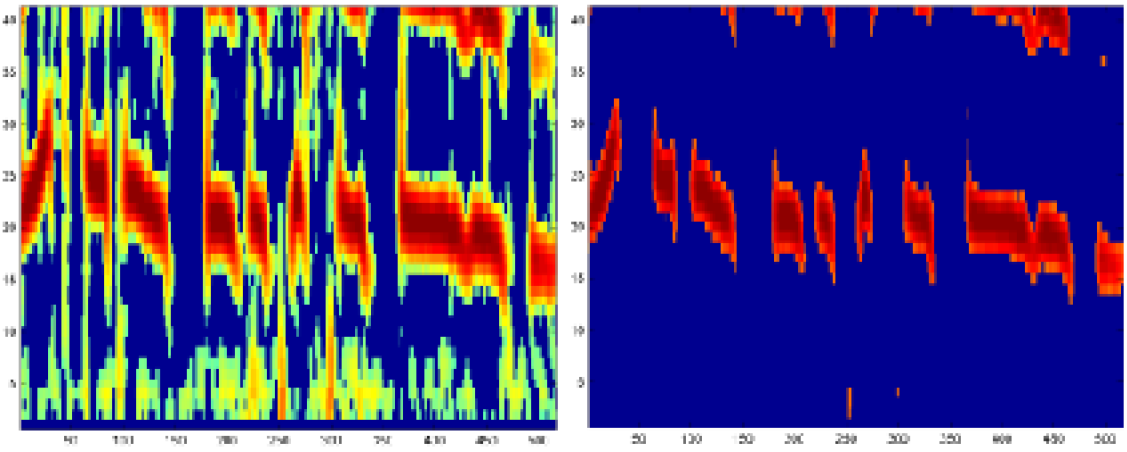


Figure 3.16 Représentation de F0 par l'algorithme Melfilter (La valeur du seuil est de 30 à gauche et 50 à droite).

Lorsqu'un seuil de 50 est appliqué nous retrouvons les performances rendant compte

des classes rythmiques (ANOVA : p=0.001, F=14.7). Pour un seuil supérieur à 70, la plupart des fichiers Anglais ont des valeurs nulles. Dans ce cas, l'Anglais est discriminé des deux autres langues, parce qu'il n'est représenté par aucune valeur.

Nous retrouvons des résultats obtenus par Ramus (1999) avec des sujets adultes, qui n'ont pas de connaissances a priori sur ces langues. Ceux-ci ne distinguent pas l'Anglais du Japonais à partir de l'intonation seule (condition 'aaaa' 51 %). Cependant, notre représentation de F0, obtenue à partir de l'autocorrélation, contient quelques informations rythmiques, puisque la courbe de F0 n'est pas interpolée.

Pour conclure sur l'origine des indices permettant d'effectuer les discriminations de l'Anglais et du Japonais, il faudrait pouvoir utiliser la synthèse 'sasasa' de façon à contrôler tous les paramètres. Les conditions d'enregistrement influent probablement sur l'intensité des fichiers. Par exemple, les fichiers Japonais pourraient avoir un volume sonore légèrement plus fort que les fichiers des deux autres langues. Néanmoins, si une distinction s'opère par le rythme, elle doit s'appuyer sur des différences d'intensité des basses fréquences du signal.

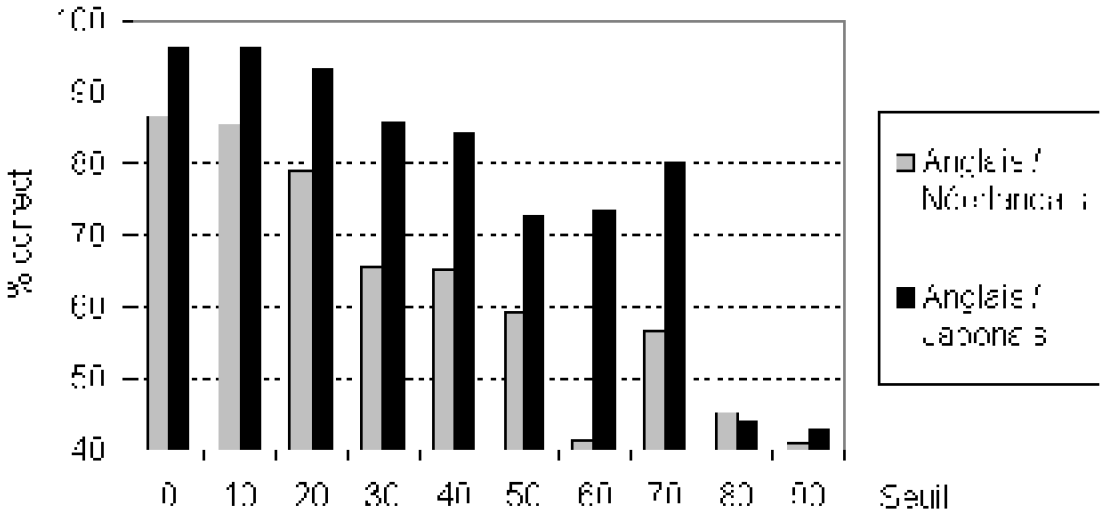


Figure 3.17 Taux de discrimination en fonction du seuil appliqué à la représentation donnée par l'algorithme Melfilter [89] *.*

V.Discussion

Cette discussion s'orientera suivant quatre points :

[89] *Les performances peuvent atteindre 41 %, qui est le pourcentage de phrases Anglaises. Dans le cas où les prototypes des deux langues sont confondus, le réseau répond toujours Anglais. Le même graphique devrait pouvoir être obtenu avec le spectrogramme à bande étroite, en modifiant le coefficient multiplicatif.*

1. Un bilan des expérimentations sera conduit afin d'indiquer le choix des représentations du signal et des méthodes de traitement ;
2. Les performances du réseau TRN seront comparées avec les travaux existants pour l'IAL à partir de la prosodie ;
3. Les perspectives découlant de ces expériences ;
4. Le rythme des langues sera abordé avec la discussion des expériences de discrimination de langues de différentes classes rythmiques.

V.1.Résumé des expérimentations d'IAL

Le premier but de ce chapitre était de proposer un modèle susceptible de traiter la prosodie de manière globale pour l'identification automatique des langues. Appliquer une tâche d'IAL au réseau TRN a permis d'élargir dans un premier temps le nombre de dimensions acoustiques traitées par le réseau : le rythme donné par les consonnes et voyelles (Pellegrino, 1998 ; Pellegrino et coll., 2002), puis l'intonation, l'intensité et les formants (représentation spatio-temporelle du signal : cochléogramme, spectrogramme, etc.).

V.1.1.Le rythme

Le but de cette première section était de retrouver les capacités du réseau pour traiter des séquences de consonnes / voyelles. Le comportement du réseau TRN a été étudié face à des séquences plus longues et plus complexes que celles auxquelles il avait déjà été testé (Dominey et coll., 1995). Ainsi, le réseau a été appliqué à des séquences représentant 5s de signal de parole et non plus des séquences de 25 éléments abstraits.

Dominey et Ramus (2000) avaient déjà proposé de simuler le comportement des nourrissons pour la discrimination des langues de classes rythmiques différentes. Nous avons montré qu'il fallait modifier ces mêmes méthodes, en apportant une mémoire supplémentaire (cf. IV.2), pour pouvoir identifier cinq langues européennes à partir d'une segmentation automatique des consonnes et des voyelles. Ainsi, en apportant une mémoire auxiliaire au réseau, il est possible d'arriver à 50 % d'identification correcte. Les performances sont augmentées lorsqu'une procédure de validation croisée est mise en place (65% en 20s). Par conséquent, les performances du réseau TRN sont améliorées lorsque le nombre de données disponibles en apprentissage est augmenté. Cependant, ces résultats sont obtenus avec le réseau sélectionné pour donner le meilleur taux d'identification en validation. Il semble difficile d'obtenir un modèle plus précis de la caractérisation globale du rythme d'une langue.

Après avoir testé un certain nombre de méthode pour améliorer les performances du réseau TRN, nous proposons de compléter l'information prosodique dirigée vers le réseau.

V.1.2.Représentation spectrographique non segmentée

Le cochléogramme permet d'obtenir des détails temporels fins (jusqu'à 30 ms) et traiter toutes les dimensions prosodiques en même temps. Cet encodage inclut la prosodie (F0 et intensité) et le spectre dans une représentation commune en fonction du temps. En outre, les évènements sont représentés à une échelle inférieure à celle de la plus petite unité linguistique qui est le phonème. L'identification se fait alors sans segmentation du signal acoustique. Cette représentation ne cherche, en revanche, pas à isoler le signal de parole du bruit de fond.

Nos recherches ont montré que les langues du corpus MULTEXT étaient corrélées au bruit de fond. Effectivement, de très bonnes performances d'identification (60 % pour la segmentation et le cochléogramme) sont atteintes avec une seule trame représentant 30 ms de signal. Ce problème est propre au corpus MULTEXT, mais lorsque la totalité du signal acoustique est utilisée, nous ne pouvons pas garantir que seule la parole permet de distinguer les langues. Le signal acoustique a donc été coupé après 2 secondes afin d'éliminer les zones de silences qui pouvaient servir pour l'identification des langues. Se faisant, le taux d'identification est au niveau du hasard au début du signal étudié.

V.2.Comparaison des performances

Pellegrino et coll. (2002) ont également proposé des résultats pour l'IAL sur le corpus MULTEXT, en utilisant une technique de validation croisée, pour une représentation incluant le rythme des consonnes et des voyelles (79 % en 20s ; > 60% pour 5s). En tenant compte d'une modélisation acoustique des voyelles, les performances atteignent 69 % en 5 secondes (Farinas, 2002). L'approche qu'ils retiennent est basée sur la description de pseudo-syllabe. En outre, ils emploient un détecteur automatique des voyelles. Nos performances sont donc comparables avec celles-ci pour 5 secondes de signal (63 %), alors que nous faisons appel à une caractérisation globale du signal acoustique sans segmentation.

Une des difficultés d'une application d'IAL est de passer d'un discours articulé pour des conditions de laboratoire à un signal téléphonique comme celui du corpus OGI-MLTS. Par exemple, pour la distinction entre le Français et l'Anglais les performances diminuent de 99 % à 76 % pour 2 secondes de signal de parole dans l'étude de Lamel et Gauvain (1994). La multiplicité des locuteurs peut être à l'origine de cette diminution de l'identification. Effectivement, le corpus OGI-MLTS contient une centaine de locuteurs par langues. En outre, des études perceptuelles ont confirmé les difficultés des adultes face à des langues inconnues, lorsqu'ils sont testés avec un nombre important de locuteur (Bond et coll., 1998).

L'application de nos méthodes à ce corpus est aussi problématique. Elles se révèlent inefficaces lorsque six langues doivent être identifiées, cependant les performances pour deux paires de langues du corpus OGI-MLTS avec l'Anglais sont nettement supérieures au hasard. En outre, les performances d'identification augmentent peu au cours du temps, suggérant la présence d'un indice tout comme pour le corpus MULTEXT. Mais ce profil de performance n'apparaît pas pour la paire de langues Allemand – Hindi. Il est ainsi vraisemblable que l'Anglais se différencie de toutes les autres langues sélectionnées. Cependant, ce problème n'a pas été soulevé par d'autres études effectuées sur le corpus

OGI-MLTS.

V.3.Perspectives pour l'IAL et la prosodie

L'identification Automatique des Langues nécessite encore des recherches approfondies pour tirer parti des informations prosodiques. Effectivement, notre étude conjointement aux études précédentes (Cummins et coll., 1999 ; Thymé-Gobbel et Hutchins, 1996 ; Itahashi et Du, 1995 ; Itahashi et coll., 1999 ; Pellegrino et coll., 2002) montrent l'écart de performances [90] entre les systèmes IAL classiques et les systèmes ayant recours à la prosodie. Il est particulièrement difficile d'obtenir un modèle global satisfaisant de la prosodie d'une langue, en particulier pour l'intonation d'une langue.

Cependant, la prosodie apparaît comme un indice potentiel au travers d'une étude perceptuelle effectuées avec des sujets adultes connaissant au moins une des langues à identifier (Muthusamy et coll., 1994). Lorsque les locuteurs ne connaissent pas la langue, seule la discrimination des langues est abordée (Stockmal et coll., 1996 ; Stockmal et Bond, 1998 ; Bond et coll., 1998 ; Lorch et Meara, 1989 et 1995). Ces articles soulignent alors les difficultés exprimées par les sujets.

En appliquant une technique de synthèse de parole, Ramus et coll. (1999) ont montré que des auditeurs ayant pour langue maternelle le Français pouvaient effectuer la distinction entre le Japonais et l'Anglais à partir de la seule succession des consonnes et des voyelles. Il conviendrait de tester si cette distinction est mieux perçue par les sujets natifs du Japonais et de l'Anglais, afin de savoir si ils ont une meilleure représentation de la prosodie de leurs langues.

Effectivement il est possible que les modèles employés en IAL pour décrire la prosodie soit incomplets. D'une part, il devrait y avoir plusieurs modèles prosodiques, et non un seul par langues. D'autre part, il devrait être établi en fonction d'autres unités linguistiques comme les phrases ou les mots. Les frontières des phrases et des propositions pourraient être retrouvées par l'intermédiaire de la prosodie à partir des pauses ou de la déclinaison de l'intonation. Ceci a été compris par les linguistiques étudiant la prosodie (Cutler, 1996), mais ils mettent en avant aussi les différences existant entre les langues (Hirst et Di Cristo, 1998). Ces différences prosodiques entre les langues sont le plus souvent définies en fonction du contenu d'une phrase.

La position de l'accent par rapport aux mots permet de distinguer les langues à accents fixes, et à accents variables. Ainsi, l'Anglais est caractérisé par un rythme trochaïque, alors que le rythme du Français est ïambique. Il est également possible de tenir compte des structures syntaxiques pour qualifier la prosodie d'une langue. Ainsi, en tenant compte de la déclinaison de la F0 et des pauses de longue durée, la structure en proposition peut être obtenue, la position des accents dans ces structures pourrait permettre d'identifier une langue.

La prosodie peut se révéler utile pour l'IAL en tant que guide afin de restreindre le

[90] En terme de pourcentage d'identification, du nombre de langues testées, des conditions d'enregistrement des corpus (laboratoire ou spontanée, lignes téléphoniques).

nombre de langues à étudier (Leavers et Burley, 2001). Dans ce contexte, la prosodie pourrait être employée pour choisir le décodeur phonétique le mieux adapté à une langue.

V.4.Simulation de la discrimination des langues en fonction des classes rythmiques

Nos expériences consacrées à l'Identification Automatique des Langues ont montré les difficultés du traitement d'un nombre élevé de langues. La technique employée repose sur des connaissances acoustiques du signal. La modélisation proposée s'apparente plus aux mécanismes utilisés par les nourrissons, qui sont influencés par la voix du locuteur, et des indices prosodiques comme le rythme ou la tonalité (Bond et coll., 1998).

Dans le même temps, nous avons montré qu'il est possible d'utiliser une représentation spectrographique sans segmentation avec le réseau TRN pour effectuer cette identification. Il est intéressant de vérifier que cette représentation peut rendre compte des propriétés rythmiques des langues, comme la segmentation en consonnes et voyelles (Dominey et Ramus, 2000) et donc fournir un modèle de réseau plus complet pour le traitement de la parole, puisque le signal de parole est traité directement. Nous avons donc effectué les simulations TRN avec trois représentations spectro-temporelles de la fréquence fondamentale. Seules les représentations tenant compte de l'intensité permettent de distinguer l'Anglais du Japonais, qui appartiennent à deux classes rythmiques distinctes.

Nous avons montré que pour pouvoir retrouver les profils de performances des nourrissons (Nazzi et coll., 1998) il fallait éliminer les valeurs d'intensité les plus faibles de la représentation spectro-temporelle de la fréquence fondamentale.

Ainsi des propriétés acoustiques, sans segmentation phonétique, permettent d'identifier des langues uniquement lorsqu'elles appartiennent à des classes rythmiques distinctes. Le fait que le réseau TRN soit influencé par les classes rythmiques des langues à discriminer suggère que le mécanisme de traitement global que nous proposons peut refléter les propriétés du traitement observé chez les enfants (Nazzi et coll., 1998).

Il faut aussi remarquer que même si nous retrouvons le même type de performance, le mécanisme que nous proposons peut être différent de celui présent chez le nourrisson, ou chez les singes. Cependant, le modèle que nous utilisons est initialement inspiré de l'architecture fronto-striatale du singe (Dominey et coll., 1995), et simule l'apprentissage de séquences sensori-motrices. Ceci renforce l'idée qu'un tel mécanisme puisse être également présent chez le singe pour traiter des séquences sonores.

De nombreux articles ont démontré qu'il était possible de retrouver ces classes rythmiques. Néanmoins, ils font appel à une segmentation du signal en consonnes / voyelles effectuée soit à la main (Grabe et Low, 2002 ; Ramus, 1999), soit à partir d'une segmentation automatique (Pellegrino et coll., 2002 ; Galvès et coll., 2002). Le réseau récurrent temporel ne nécessite pas de segmentation explicite en unité proche des phonèmes.

Nos expériences menées en discrimination suggèrent que l'intensité des différentes

bandes de fréquences distingue les langues. Cependant, il reste possible que cela soit lié aux conditions d'enregistrement (comme la distance entre le locuteur et le micro). Afin de vérifier que ce sont bien les propriétés rythmiques des langues qui sont à l'origine du profil de discrimination, les réseaux TRN devraient être testés avec juste les phrases synthétisées en stimuli « sasasa » (Ramus, 1999). En outre, l'utilisation de carte de Kohonen (1982) pour classer les motifs fournis par le réseau TRN permettrait de simuler un apprentissage non supervisé. Il ne resterait plus qu'à simuler le changement de succion pour répliquer les expériences effectuées avec les nouveau-nés (Nazzi et coll., 1998 ; Ramus, 2002b), avant de tester notre modèle avec d'autres langues.

VI.Conclusion

Nous montrons donc que le réseau récurrent temporel peut extraire les régularités acoustiques et prosodiques au cours du temps, pour identifier automatiquement les langues. Dans un premier temps, le réseau TRN a pu traiter des structures temporelles définies sur plusieurs dimensions prosodiques, s'étendant sur des passages de plusieurs secondes. Au cours de ces expériences, nous avons montré que le réseau TRN traite une représentation de plus en complète du signal : rythme, fréquence fondamentale, et formants sont traités dans une représentation commune et avec une grande précision temporelle (Cochléogramme).

Dans un second temps, le réseau TRN a traité la structure prosodique de phrases de langues appartenant à différentes classes rythmiques, montrant ainsi que le réseau TRN était toujours sensibles aux différences rythmiques, même lorsque les consonnes et voyelles ne sont pas désignées au sein du signal. Le réseau TRN distingue des langues uniquement lorsque leurs classes rythmiques différent, comme c'est le cas chez les nourrissons (Nazzi et coll., 1998).

Berkling (1996) précise que l'Identification Automatique des Langues se fonde sur une représentation sous forme d'unités discrètes de la parole. Nos premières expériences suivent ce principe, puisque le signal de parole est réduit à des segments fournis par la distinction entre consonnes et voyelles (Pellegrino, 1998). Nos dernières expérimentations montrent cependant qu'il est possible de ne pas chercher d'unité particulière, dans le signal de parole (*cf.* section III.4. Simulation de discrimination de langues).

L'objet de ce premier chapitre expérimental était de démontrer que le réseau TRN pouvait être employé pour identifier 5 langues pour une fenêtre temporelle longue, de l'ordre de plusieurs secondes, tout en étant influencé par le rythme propre d'une langue. Le premier point de l'hypothèse de Continuum Temporel est donc vérifié.

Est-il possible d'employer le réseau TRN pour identifier différentes attitudes prosodiques, exprimées dans une même langue ? Ainsi, ces expériences nous permettront de tester la validité du réseau TRN pour traiter différents contours de la fréquence fondamentale. Dans ce contexte également, les informations transmises seront

uniquement acoustiques, de la même manière qu'un nourrisson appréhende les différentes expressions, que lui communiquent ces parents.

Chapitre Quatre Thème 2 : Identification Automatique des Attitudes Prosodiques

J'ai souvent été frappé comme d'un fait très curieux, de ce qu'un si grand nombre de nuances d'expressions soient reconnues instantanément, sans que nous ayons la conscience d'un effort d'analyse de notre part. je ne crois pas que personne puisse décrire nettement une expression maussade ou maligne : cependant des observateurs en grand nombre, déclarent unanimement que ces expressions sont reconnaissables chez les diverses races humaines. […] Il en est de même d'un grand nombre d'autres expressions, qui m'ont fourni l'occasion d'éprouver combien il faut se donner de peine pour montrer aux autres quels sont les points qu'il faut observer. Traduit de Darwin (1872). ; Rimé et Scherer (1989, p. 175). Les émotions. Textes de base en psychologie.

I.Introduction

Le chapitre précédent a démontré que le TRN pouvait traiter la structure temporelle des langues, afin de les identifier, apportant ainsi une première validation de l'hypothèse de Continuum Temporel pour une organisation globale des indices prosodiques. Dans ce

chapitre, nous continuons cette validation, en abordant une nouvelle relation entre la structure temporelle et l'organisation fonctionnelle du langage. Les résultats expérimentaux de l'identification des attitudes prosodiques chez l'être humain (Morlec, 1997 ; Morlec, Bailly et Aubergé, 2001) ont révélé que les attitudes prosodiques pouvaient être distinguées avec un score de 72,9 %. L'objectif des expériences suivantes est de tester si le réseau récurrent temporel peut catégoriser des attitudes prosodiques à partir du contour de la fréquence fondamentale en respectant l'hypothèse de contrainte temporelle.

Le réseau TRN sera testé avec un corpus initialement développé pour créer un système de synthèse de la fréquence fondamentale, pour certaines attitudes prosodiques données (Morlec, 1997 ; Morlec et coll., 2001). En premier lieu, un état des lieux de la reconnaissance des émotions, thème assez proche des attitudes prosodiques, sera dressé, en considérant les techniques d'identification automatique, et une étude comportementale d'identification des attitudes prosodiques en Français. Le matériel sur lequel repose cette étude sera alors présenté, avant de détailler les simulations avec le modèle TRN. Enfin, nous discuterons des résultats obtenus, en les comparant à des études antérieures.

II.Contexte de l'identification des attitudes prosodiques

Lors de l'utilisation de machine pour traiter la parole, les attitudes du locuteur peuvent être mises à contribution pour mieux saisir ses intentions. En outre, dans le cas de traduction automatique, il serait intéressant de pouvoir transmettre simultanément les émotions ou l'attitude du locuteur, de façon à ce que l'interlocuteur puisse comprendre au mieux ses désirs, plus particulièrement dans le cas où ils ne se trouvent pas face à face (ligne téléphonique). Cowie et coll. (1991) ont suggéré que les émotions se définissent à partir d'une combinaison d'émotions simples, à l'image des couleurs qui peuvent toutes être définies en fonction de trois couleurs, dites primaires. Les émotions de bases le plus souvent retenues incluent la joie, la tristesse, la peur, la colère, la surprise et le dégoût.

Cette section présente deux types de travaux : ceux concernant l'identification automatique des émotions, puis deux études expérimentales s'appuyant sur le même matériel que celui avec lequel le TRN a été testé.

II.1.Reconnaissance automatique des émotions

L'identification des émotions ou des attitudes prosodiques est un thème de recherche assez récent. Nous allons dresser un panorama des diverses techniques employées pour executer cette tâche. Pour les études existantes, le nombre de données est relativement restreint, en terme de quantité d'exemples appris, de paramètres caractérisant le signal, et de méthodes employées (McGilloway et coll., 2000; Breazeal, 2000 ; Slaney et Mc

Roberts, 1998). Une des premières études sur les émotions portait sur les corrélats acoustiques des attitudes. Ainsi l'expression de la joie a une fréquence fondamentale moyenne plus élevée que des phrases calmes.

Contrairement à l'Identification Automatique des Langues, il n'existe pas de corpus commun. La plupart des études existantes effectuent apprentissage et validation avec un seul locuteur. ASSESS (McGilloway et coll., 2000) permet des performances de 55 % pour 4 attitudes basiques à l'aide d'une analyse discriminante. Dellaert, Polzin et Waibel (1996) proposent un système utilisant la fréquence fondamentale et un algorithme des *n* plus proches voisin, avec des performances de l'ordre de 80 % pour 4 émotions. Slaney et Mc Roberts (1998 ; 2003) ont axé leur recherche sur l'identification des émotions chez les enfants.

McGilloway et coll. (2000) prennent en compte plusieurs paramètres statistiques : la moyenne, minimum et maximum, la différence entre les extrêmums, la variance et la distribution de l'intensité, la longueur des segments syllabiques ou phonémiques, et les montées de F0. Ils ont employées des « support vector machine », des mixtures gaussiennes et des analyses discriminantes linéaires.

Oudeyer (2002) a étendu ces travaux, en prenant en compte la fréquence fondamentale, ainsi que l'intensité des bandes de fréquences les plus basses (< 250 Hz), et des bandes spectrales plus hautes (>250 Hz). Une mesure spectrale est également incluse, calculée à partir du vecteur absolu dérivé des 10 premiers coefficients MFCC (Mel Frequency Cepstral Coefficients). Chaque mesure est effectuée toutes les 10 ms, et est traduite en quatre séries : les valeurs brutes, les minima, les maxima, les durées entre deux extrêma. Chaque série est alors caractérisée par la moyenne, le minimum, le maximum, la différence entre le minimum et le maximum, la variance, la médiane, le premier et troisième quartile, l'interquartile et la moyenne absolue de la dérivée locale. Ceci conduit donc à 5 dimensions, représentées par 4 séries décrites par 10 statistiques, soit un total de 200 paramètres. Ces paramètres sont normalisés, avant d'être appris. La base étudiée est composée de 200 exemples par locuteur et par émotion, soit 2000 exemples au total. Il s'agit de courtes phrases, comme « Bonjour », « ça va », « Qu'est ce que vous aimez mangez ? ».

A l'aide d'une technique de validation croisée avec 90 % du corpus en apprentissage et l'ensemble des paramètres, quatre émotions sont reconnues avec un taux d'identification situé entre 92 % et 97 %. Les statistiques décrivant l'intensité de la partie prosodique du signal sont souvent prises en compte dans les règles élaborées par les arbres de décision. Une mesure de l'entropie des différents paramètres permet de sortir les 20 paramètres les plus informatifs pour l'identification des émotions. Parmi ceux-ci trois seulement sont cités dans des études psychoacoustiques (la moyenne, le minimum et le maximum de la hauteur).

Nwe, Foo et De Silva (2003) ont testé leur système avec deux langues, le Mandarin et le Birman. Le corpus a été validé par une identification des émotions par des êtres humains (performance de 65.7 %). Des chaînes de Markov cachés sont employées pour traiter l'information venant du signal de parole. La base d'apprentissage est composée de 60 % des phrases de chaque locuteur. L'algorithme proposé est dépendant du locuteur,

mais indépendant du texte. L'utilisation des coefficients LFPC et des chaînes de Markov Cachées (HMM) permet d'identifier 6 émotions différentes avec un score de 80 %, et ce, avec 2 langues différentes. En outre, ils obtiennent les scores les plus élevés avec les coefficients LFPC (Log Frequency Power Coefficients), qui sont comparés aux coefficients traditionnels utilisés en traitement de la parole MFCC, et LPCC (Linear Prediction Cepstral Coefficients). Les coefficients LFPC conservent mieux les valeurs de la F0 pour les filtres basses fréquences.

Une seule étude propose des résultats avec plusieurs locuteurs (50 pour Nicholson, Takahashi et Nakatsu, 1999 et 2000). Un ensemble de motifs acoustiques (phonétique et prosodique) est calculé avant d'être transmis à un réseau de neurones. Les indices sont prosodiques : l'énergie du signal et la hauteur, et phonétiques : 12 paramètres LPC (Linear Prediction Coefficients) et un paramètre ΔLPC de variation de ces paramètres. Chaque phrase est représentée par un ensemble de 20 vecteurs de 15 indices. Un vecteur de 300 composantes est donc donné en entrée de huit réseaux, qui apprennent chacun une émotion différente. Chaque réseau donne alors un indice de vraisemblance pour chaque émotion apprise. Chaque réseau comprend 4 couches, la dernière ne comprend qu'un seul neurone. L'apprentissage est effectué par une rétropropagation du gradient. 50 hommes et 50 femmes constituent la base de données. Les hommes et les femmes sont séparés pour l'apprentissage. Un locuteur ne se retrouve pas à la fois dans le corpus d'apprentissage et de test. Les performances sont environ de 50 % lors du test pour les hommes et les femmes, et ceux avec la combinaison de réseaux, mais également avec un seul réseau de taille plus importante.

Quelles sont les capacités de perception des attitudes prosodiques par les être humains en Français ?

II.2.Expérimentation chez l'être humain

Les expériences conduites chez l'être humain ont été effectuées à partir du corpus utilisé dans cette étude. Les êtres humains ont eu à leur disposition des stimuli de parole naturelle, où plusieurs dimensions prosodiques sont présentes.

II.2.1.Reconnaissance des attitudes (Aubergé et coll., 1997)

Six auditeurs de langue maternelle française ont réalisé cette expérience. Ils n'ont pas reçu d'entraînement préalable au test. Après l'écoute d'un stimulus, chaque auditeur doit effectuer un choix parmi l'une des six attitudes (choix ferme et forcé). Les 19 phrases du corpus sont présentées deux fois dans un ordre aléatoire sans que deux attitudes identiques ou deux phrases identiques ne se succèdent. Le corpus compte une phrase de une syllabe, trois de deux syllabes, quatre de trois syllabes, cinq de quatre syllabes, et six de cinq syllabes. Avec les 6 attitudes cela représente 228 stimuli.

Le taux de reconnaissance est assez élevé, environ 72.8 %. Les deux catégories les mieux discriminées sont les attitudes modales (Déclaration et Question). Ces tests montrent que nous sommes capables de véhiculer des attitudes (positionnement du locuteur par rapport à son propre discours, phrases méta-discursives) par la prosodie

seule, les phrases porteuses n'étant ni marquées lexicalement ni syntaxiquement.

II.2.2.Dévoilement progressif (« gating ») (Aubergé et coll., 1997)

Les données apportées par ce type d'expérience permettent d'évaluer quelle peut être la quantité d'information nécessaire pour distinguer une attitude prosodique.

Le paradigme du dévoilement progressif consiste à dévoiler progressivement un stimulus de parole, depuis son commencement jusqu'à sa fin. Pour cette étude, les stimuli sont des phrases de 5 syllabes. Le découpage du stimulus est temporel et s'effectue syllabe par syllabe. Cette portion du signal est nommée fenêtre (« *gate* »).

Dans ces travaux sur le Danois, Thorsen (1980) dévoile progressivement des segments de phrases non typées provenant de trois types d'énoncés : déclaratif, continuatif et interrogatif. Les sujets ont ainsi à deviner qu'elle est la fonction de l'énoncé dont ils n'entendent qu'une partie. Les résultats montrent que le taux d'erreur décroît lentement jusqu'au premier groupe accentuel pour lequel il chute rapidement. Ainsi, elle montre que le début de la phrase transmet assez d'éléments à même d'autoriser une reconnaissance correcte. Le Hollandais Van Heuven et ses collaborateurs (1997) ont également réalisé une expérience de dévoilement progressif sur deux attitudes en Néerlandais. Ils concluent aussi que les sujets sont capables d'identifier ces deux attitudes prosodiques, bien avant la fin de la phrase.

Cette expérience fait appel uniquement aux 6 phrases de 5 syllabes du corpus « identification et gating » (Morlec, 1997). Le découpage est syllabique, les syllabes ôtées sont substituées par un bruit blanc dont la durée a été fixée à deux secondes. Cette mesure interdit à l'auditeur d'estimer la longueur totale de la phrase entendue. Les six attitudes sont représentées pour chaque phrase, soit 180 stimuli pour un auditeur.

Six sujets, de langue maternelle française, ont participé à l'expérience de dévoilement progressif. Ils n'ont reçu aucune phase d'entraînement. Pendant l'expérience, l'auditeur garde la trace écrite de la phrase qu'il a traitée, afin de minimaliser le nombre d'ambiguïté liée à la sémantique. Il est également informé de la définition des attitudes.

Dés la deuxième syllabe, le taux de reconnaissance est élevé. De plus, la plus grande progression se situe entre les deux premières syllabes (+27 %), ce qui corrobore les études antérieures. Ces tests montrent que les attitudes prosodiques sont accessibles à l'auditeur de manière robuste et précoce, avant la fin de l'énoncé (voir Figure 4.4).

Les travaux présentés ci-avant prouvent que les émotions sont portées par plusieurs composantes prosodiques et phonétiques. Cependant, notre objectif au cours de ce chapitre est de montrer que le réseau TRN peut encoder non seulement la structure rythmique, mais aussi la structure donnée par le contour de la fréquence fondamentale. C'est pourquoi nous ne retiendrons que ces composantes (F0 au cours du temps) pour décrire les différentes attitudes.

III.Matériel et méthodes

Cette section décrit les attitudes prosodiques, ainsi que le corpus dont elles sont issues.

III.1.Les attitudes prosodiques

Cette étude est soumise à certaines contraintes. La première d'entre elles est de ne retenir que la fréquence fondamentale pour valider son traitement par le TRN. Enfin la seconde est d'employer un corpus d'attitudes prosodiques, défini pour créer un système de synthèse de la prosodie pour les composantes du rythme et de la fréquence fondamentale. A cet effet, un ensemble de six attitudes avait été conçu. Elles peuvent se distinguer en deux groupes (qui ne sont pas pris en compte dans notre travail) :

- Modalités. Le locuteur s'efforce de ne pas faire apparaître ses sentiments dans ses énoncés :
  - Déclaration (DC);
  - Question simple (QS) ;
- Intonations de discours ; Ces quatre attitudes informent sur la position de l'auteur face à ses propos :
  - Exclamation de surprise (EX) : stupéfaction du locuteur ;
  - Doute-incrédulité (DI) : désaccord partiel avec ce qui a été exprimé précédemment;
  - Ironie de soupçon (SC) : doute sur l'affirmation de l'interlocuteur ;
  - Evidence (EV) : croyance profonde du locuteur en ses dires.

Le travail original avait choisi d'étudier dans le même temps les fonctions syntaxiques et expressives que peuvent véhiculer les attitudes prosodiques. Dans ce contexte, nous chercherons pas non plus à distinguer ces deux fonctions qui ne semblent pas différentiables dans la pratique (Morlec, 1997). Dans le cadre de cet étude nous limiterons notre approche au Français tout en remarquant que les émotions ont un caractère universel, qui font qu'elles peuvent être reconnues indépendamment de la langue, tandis que les modalités sont définies de façon différente pour chaque langue. Cette remarque n'intervient pas dans notre travail, dans la mesure où seul le Français sera considéré.

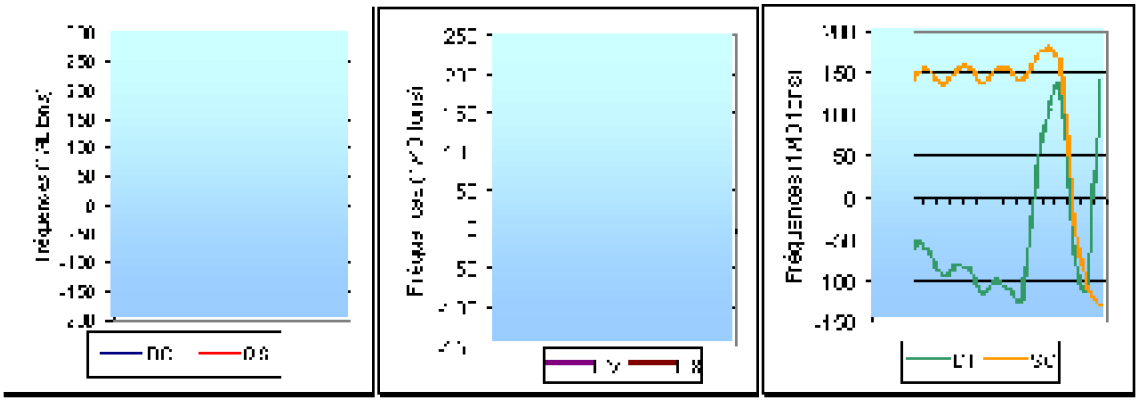


Figure 4.1 Variation de F0 pour les 6 attitudes prosodiques (moyenne des phrases du corpus) DC : Déclaration ; QS : Question simple ; EX : Exclamation de surprise ; DI : Doute-incrédulité ; SC : Ironie de soupçon ; EV : Evidence.

Comme suggéré par les prototypes moyens, les différentes attitudes se distinguent toutes par leur contour intonatif (Figure 4.1). Effectivement, des méthodes de classification par modèle gaussien et par réseaux SRN ont retrouvé une classification quasi-parfaite (> 97%, travail de DEA en collaboration avec G. Bailly). Dans ce contexte initiale de synthèse de la fréquence fondamentale et du rythme, seules ces deux mêmes informations ont été retenues pour traduire les différentes attitudes. Ainsi chaque Groupe Inter P-Center sera décrit par quatre valeurs : son coefficient d'allongement ainsi que trois valeurs de F0 situées à 10%, 50% et 90% de la durée de ce GIPC [91] .

III.2.Le corpus retenu

Ces ensembles ont été choisis parmi le corpus tiré de la thèse de Y. Morlec (1997). Pour tester les capacités du réseau, il ne sera retenu que les phrases de 6 syllabes, qui seront divisées en deux corpus de 60 phrases chacun. Les phrases de 6 syllabes représentent le plus grand ensemble de phrases disponibles. Il est à noter que les phrases ont subit un prétraitement qui a permis de mettre en exergue quelques-uns des paramètres issus du signal de parole (phonèmes, durée, fréquence fondamentale, ratio GIPC, orthographe, syllabe, etc.). De tous ces paramètres, il ne sera garder que la fréquence fondamentale et la durée des GIPC.

III.3.Représentation de la Fréquence Fondamentale

L'encodage de la F0 pour le TRN a constitué le principal travail du DEA. Ce rapport contient un descriptif détaillé des divers types de codages testés. Nous avons testé trois méthodes de codage, en passant d'un codage discret (neurone par neurone) de la fréquence fondamentale à un codage continu en temps et en fréquence. La continuité en temps est assurée par une interpolation linéaire des données, la continuité en fréquence par une courbe de Gauss qui définit l'activité des neurones d'entrées.

[91] Les valeurs de F0 et de durée ont été fournies par les auteurs du corpus (Morlec, 1997 ; Morlec et coll., 2001).

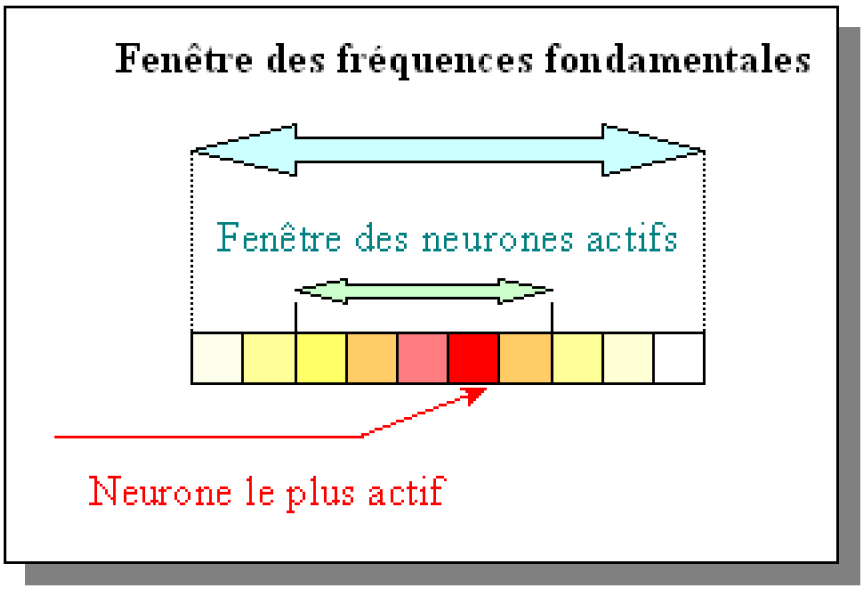


Figure 4.2 Codage continu de la F0 par une population de neurones. La taille de la fenêtre des neurones actifs dépend du paramètre sigma. Le niveau de gris indique le degré d'activation d'un neurone.

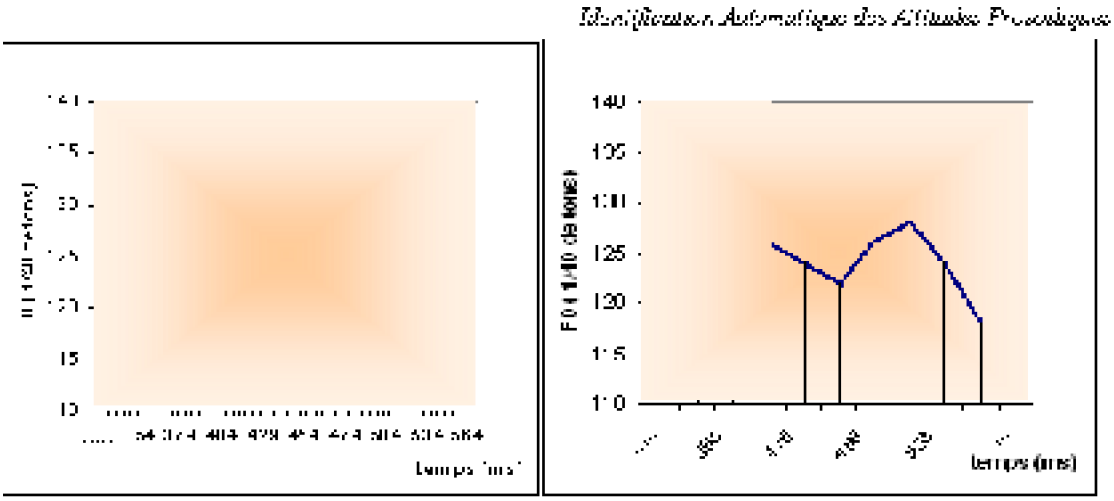


Figure 4.3 Codage de la F0 dans le temps, à gauche, valeur discrète à droite approximation linéaire dans le temps.

IV.Expérimentation

Au cours de cette section, les résultats obtenus en DEA sont brièvement énumérés. Il s'agit de valider que le réseau TRN peut traiter la fréquence fondamentale et sa structure temporelle. Ensuite deux expériences complémentaires menées durant la thèse seront décrite.

IV.1.Identification des attitudes prosodiques (Blanc et Dominey, 2003)

Le but de notre travail était de reproduire par un système informatique les résultats de tests subjectifs réalisés dans le cadre du Travail de Maîtrise de T. Grépillat (Aubergé et coll., 1997). La première partie du travail de DEA a décrit les performances de classification de modèle gaussien (99.6 % pour F0 ; 52.6 % pour le rythme) et de réseaux SRN (97 % pour F0). Les deux modèles basés sur F0 ne tiennent pas compte des durées, 3 valeurs de F0 par syllabe sont transmises au modèle. Le TRN devrait naturellement prendre sa place entre des performances humaines et les résultats des modèles mathématiques.

Le dernier type de codage donne les meilleurs résultats sur le corpus de validation, mais il fait appel au plus grand nombre de paramètres. Le nombre de neurones dont l'activité est non nulle doit se situer dans le ratio de 7 pour 15, soit un peu moins de la moitié des neurones. Cependant, l'apprentissage reste un des meilleurs sur le codage par population ne se servant que de trois neurones, c'est à dire qu'une seule unité du réseau est activée pour une plage de fréquences donnée.

D'autre part, l'utilisation de la courbe de Gauss permet l'augmentation des performances lorsque le nombre de neurones augmente de façon significative. De même, la continuité de la F0 dans le temps autorise de meilleurs résultats, pour la validation en priorité. L'apport de l'information consonne ou voyelle augmente les performances dans le cas où le nombre de neurones codant la fréquence est supérieur à celui de ceux codant consonnes et voyelles. Dans le cas contraire, cette information gêne le travail de catégorisation du réseau récurrent.

Type de codage	Nombre de neurones	Sigma	Apprentissage	Validation
non continue	3	XXXXXXXXX	79,7%	75,3%
	15	XXXXXXXXX	40,2%	35,3%
continue (gauss)	3	15	38,7%	38,7%
en fréquence	4	15	77,5%	76,6%
	10	15	67,4%	70,6%
continue en temps	3	15	66,7%	69,7%
et en fréquence	10	15	77,8%	7[illegible],4%
	1	15	36,0%	37,6%
continue	3	15	63,2%	65,1%
en temps	10	15	68,2%	70,2%
et en fréquence	15	15	72,8%	74,8%
consonnes et voyelles	15	47	7[illegible],4%	7[illegible],7%
	15	95	74,2%	77,2%
	15	70	77,2%	80,6%

Tableau 4.1 Pourcentage d'identification correcte du réseau le plus performant en apprentissage en fonction des méthodes et des paramètres pour coder F0.

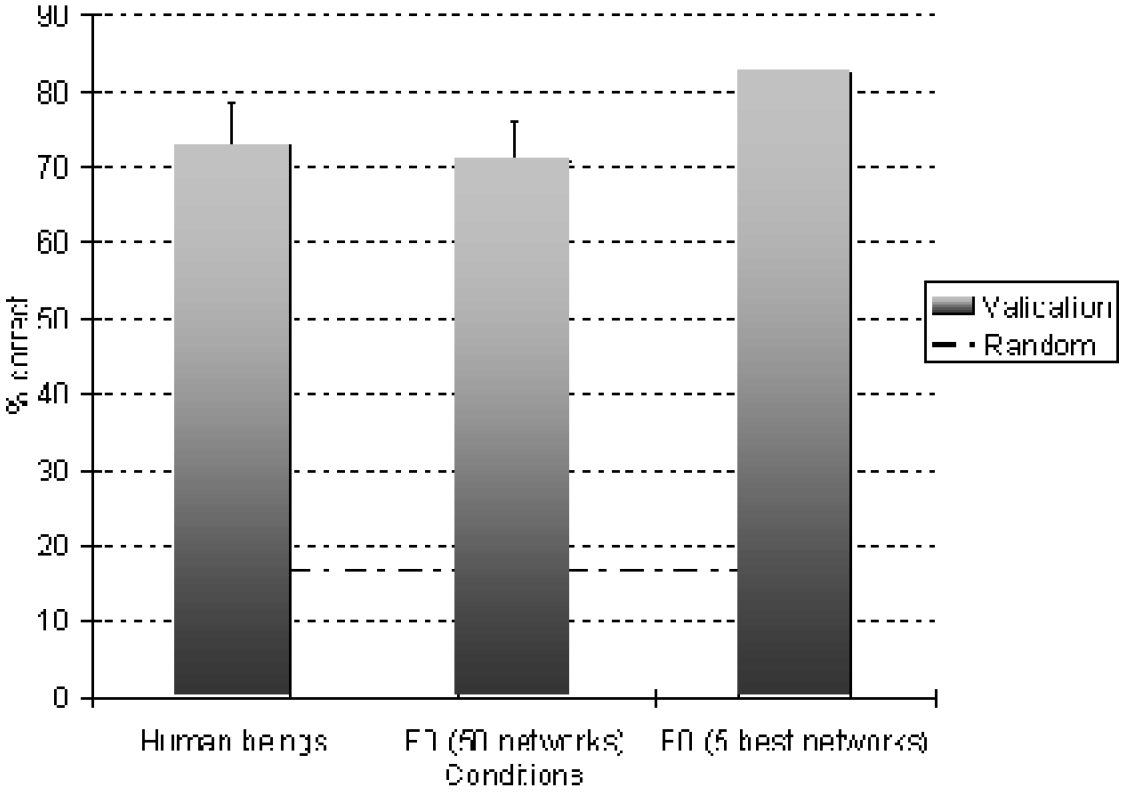


Figure 4.4 Performance d'identification des attitudes prosodiques pour les êtres humains (colonne 1), pour 50 réseaux TRN (colonne 2), et pour les 5 réseaux TRN les plus performants en apprentissage (colonne 3), tirée de Blanc et Dominey, 2003. Les barres indiquent l'écart-type des performances.

Ce dernier résultat a été repris dans un article (Blanc et Dominey, 2003), avec une population de réseaux, obtenus avec le programme du TRN réécrit en C++, et uniquement à partir de l'information de F0.

IV.2.Robustesse au ralentissement

Il serait intéressant de tester les capacités de généralisation du réseau TRN lorsque les attitudes prosodiques sont énoncés avec un tempo différent. Le meilleur moyen pour répondre à cette question est sans doute d'inclure ces phrases ralenties dans le corpus d'apprentissage. Nous avons pu tester les aptitudes du réseau en créant des séquences de F0 ralenties artificiellement par un facteur allant de 2 à 10. Les valeurs du temps sont simplement multipliées par ce coefficient. Le corpus d'apprentissage n'est pas modifié.

Bien que le ralentissement soit extrêmement exagéré, les performances restent supérieures au hasard. Les performances passent de 80 % à 60 % pour un facteur 2 de ralentissement. Malheureusement, il n'existe pas de performances humaines pour effectuer une comparaison avec le TRN.

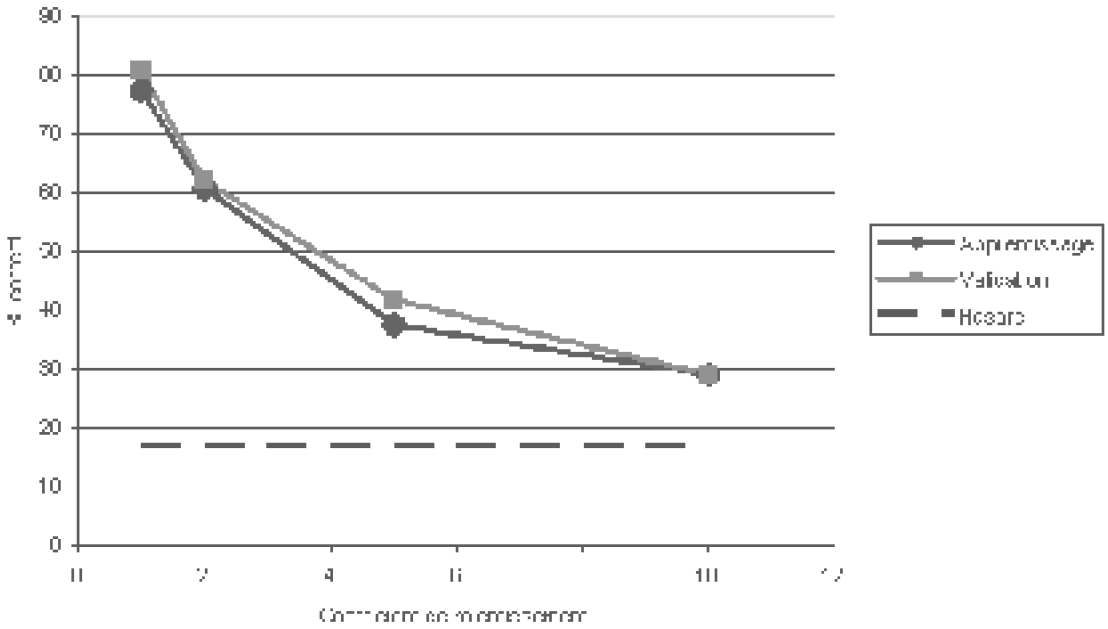


Figure 4.5 Performance du réseau TRN en fonction du facteur de ralentissement de la F0

IV.3.Méthode d'accumulation

Nous avons voulu étudier le comportement du réseau pour les méthodes choisies dans le cadre de l'IAL (Chapitre trois, III.2.4). Ainsi, l'évaluation du réseau TRN prend en compte les deux couches cachées (State et $State_D$), alors que nous ne tenions compte pour les résultats précédents que de la couche de contexte $State_D$ (Blanc et Dominey, 2003). En outre, nous comparons les deux méthodes : Accumulation utilisant l'état du réseau à chaque pas de la simulation, et la technique qui n'utilise que l'état courant du réseau (une seule trame). Ainsi, lorsque seule la trame courante du réseau TRN est employée, il convient de prendre en compte uniquement la couche de contexte : $State_D$ (63.2 % avec deux couches et avec la couche $State_D$: 70 %).

Les tests montrent que la méthode d'acccumulation donne les meilleurs résultats, et que le réseau TRN oublie certaines informations présentes dans le signal d'entrée. Toutefois, le taux d'identification reste légèrement inférieur à celui observé avec d'autres méthodes (SRN et classifieur gaussien), mais supérieur à une méthode de traitement effectuée à partir d'une moyenne simple. En revanche, l'utilisation du TRN seul donne des performances inférieures, même si seule la couche représentant le mieux les séquences est prise en compte ($State_D$, performance de 70%).

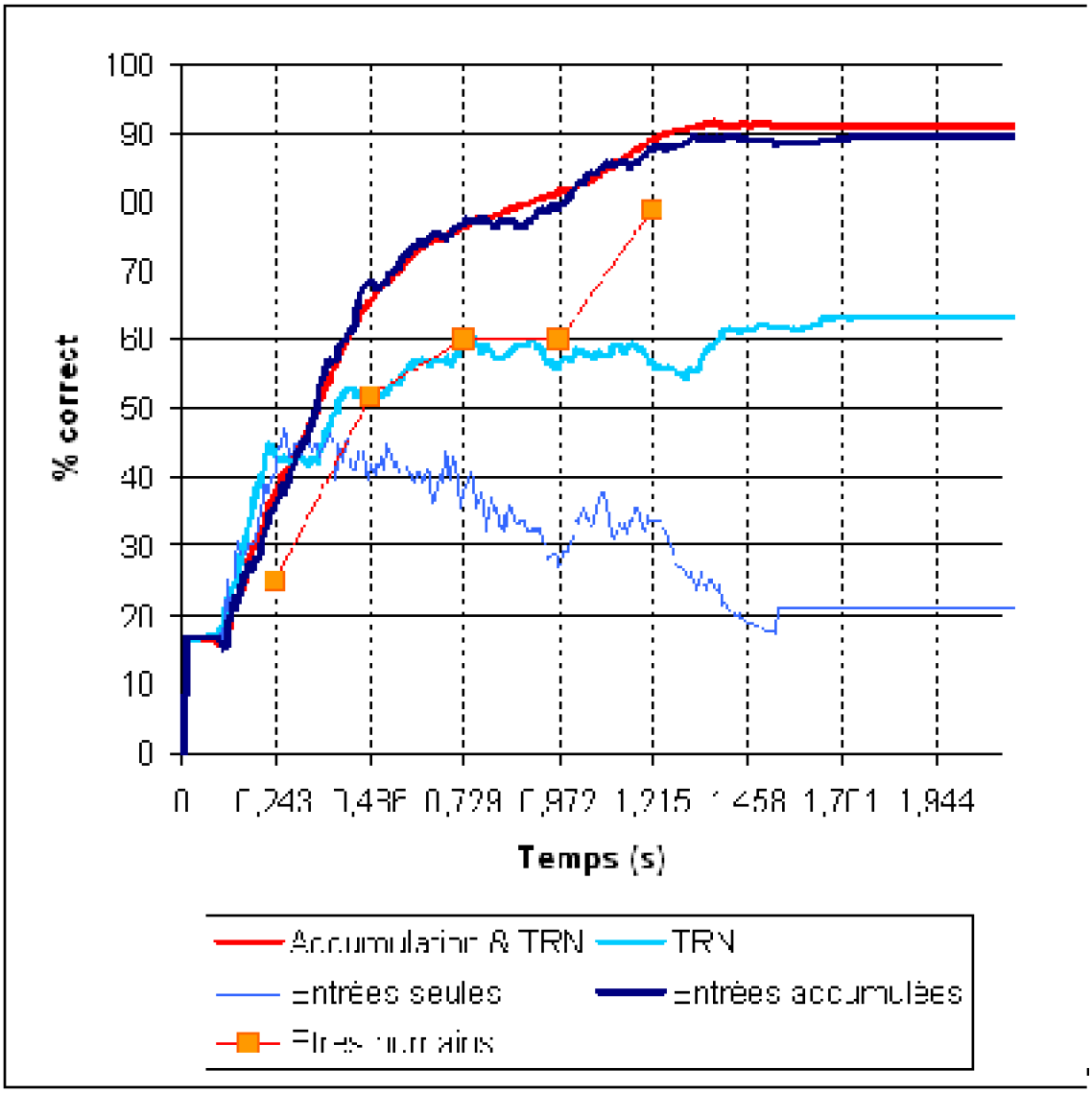


Figure 4.6 Pourcentage d'identification correcte au cours du temps des attitudes prosodiques avec la méthode d'accumulation. Les performances sont le résultat de la moyenne de la population de 50 réseaux, avec ou sans la méthode d'accumulation. Les performances sont également indiquées sans l'encodage du TRN (entrées). Les performances humaines sont celle obtenues durant l'expérience de dévoilement progressif, pour les phrases contenant 5 syllabes (Aubergé et coll., 1997).

V.Discussion

Le réseau récurrent temporel a pu être adapté à la reconnaissance des attitudes prosodiques et au traitement de la fréquence fondamentale. Ainsi ce réseau pouvait distinguer les langues à partir de leur structure temporelle donnée par les consonnes et les voyelles (Dominey et Ramus, 2000). Dans ce chapitre, la structure prosodique est

décrite par les variations de la fréquence fondamentale. Le réseau ne traite plus une suite de symboles ou d'éléments discrets (Dominey et Ramus, 2000), mais une valeur (F0) qui évolue de façon continue (pour les parties voisées du signal de parole). Ainsi, le réseau ne nécessite pas une segmentation manuelle du signal de parole. L'approche de Buhmann et coll. (2000) requiert une segmentation par syllabe pour apprendre l'intonation de 6 langues. Les approches statistiques (classifieur gaussien) et connexioniste (SRN) avaient recours à des valeurs de F0, ou de durées définies, sur chaque syllabe des phrases exprimées par le locuteur.

Le taux d'identification *in fine* est de 80 % sur l'ensemble des 6 attitudes prosodiques, en employant uniquement la trajectoire de la fréquence fondamentale. En outre, nous montrons que les performances du réseau TRN peuvent se généraliser aux variations temporelles. En effet ce réseau n'est pas trop perturbé lorsque les données qu'il a apprises sont ralenties (de 80 % à 60 % pour un facteur 2 de ralentissement). Mais nous n'avons pas de point de comparaison chez l'être humain. Effectivement, cette expérience est simulée et n'a pas été effectuée sur avec des sujets humains parlant lentement.

Une comparaison directe avec les études d'identification automatique des émotions ne peut être envisagée, dans la mesure où le matériel testé est différent (ajout de deux modalités syntaxiques, six attitudes en tout et utilisation exclusive de F0). Cependant, notre étude répond avec un taux de 80%, qui s'inscrit dans la gamme des taux d'identification variant entre 55 % (McGilloway et coll., 2000) et 97 % (Oudeyer, 2003). Toutefois, nous considérons aussi un seul locuteur. La plupart du temps, les méthodes n'ont pas été testées avec des locuteurs distincts, en apprentissage et validation. Or la variabilité entre les locuteurs est le problème le plus crucial pour le traitement de la parole. En outre, il n'existe pas de test de reconnaissance des émotions en parole continue, ce qui constituera un nouveau point de recherche pour cette thématique.

Certaines émotions ont des caractéristiques communes, ainsi la colère, la peur, la joie, et la surprise ont une amplitude et valeur moyenne plus élevées pour la fréquence fondamentale. Il est donc possible que ces caractéristiques communes altèrent les performances. Pour améliorer les performances d'identification, les attittudes prosodiques ayant des caractéristiques communes devraient être regroupées avant d'être identifiées séparément. Il est aussi probable que l'ajout d'autres dimensions à la fréquence fondamentale (comme l'intensité des basses fréquences) permettraient d'accroître les performances.

En conclusion, le réseau TRN est capable de traiter la fréquence fondamentale pour identifier différentes attitudes prosodiques, ainsi que la structure temporelle de la parole pour distinguer des langues de classes rythmiques différentes. Ces deux études nécessitent donc que le réseau TRN encode des passages de parole relativement long (supérieures à une seconde) pour pouvoir extraire les régularités et les contours permettant de répondre à une tâche de classification.

Pour répondre à notre hypothèse de Continuum Temporel, il faut que le réseau TRN puisse encoder des données plus courtes de l'ordre de mots. Ce point va être éclairci dans le chapitre suivant pour identifier les mots appartenant à des catégories syntaxiques différentes, et ce à partir de l'information acoustique englobée dans ces mots.

Chapitre Cinq Thème 3 : Identification Automatique des Mots de Fonction et de Contenu

" Le langage est comme un virus" William Burroughs
Alors comment se transmet-il ? ...

I.Le début de l'acquisition de la syntaxe : la catégorisation lexicale

Les chapitres expérimentaux précédents ont prouvé que le réseau récurrent TRN pouvait executer des tâches d'identification à partir de la prosodie pour des durées relativement longues. Il faut donc maintenant vérifier que ce même système peut être employé pour caractériser des séquences, définies sur une unité de temps plus courte : le mot. A quelle tâche d'identification faire appel pour catégoriser une structure temporelle définie sur un mot ?

Notre groupe de recherche étudie les mécanismes permettant la compréhension syntaxique. Pour effectuer cette tâche, le modèle développé (Dominey, Hoen, Blanc et

Lelekov, 2003) nécessite une classification des mots en catégories fonction et contenu. Notre objectif est donc de démontrer 1) qu'il existe au moins un indice prosodique présent dans la fréquence fondamentale permettant de distinguer les mots de fonction des mots de contenu, 2) que le modèle TRN peut traiter la fréquence fondamentale sur un mot pour réaliser cette distinction.

Pendant l'apprentissage d'une langue, les enfants doivent apprendre à classer les mots dans des catégories grammaticales appropriées. Les adultes doivent, quant à eux, effectuer ce classement rapidement et précisément. Quel peut être le type des informations qui autorisent la classification des mots en catégories grammaticales ?

I.1.Quatre origines possibles pour la catégorisation lexicale

La catégorisation lexicale est certainement l'une des questions les plus redoutables du problème de l'acquisition du langage. Plusieurs théories d'amorçages ont été proposées : l'information distributionnelle, l'amorçage syntaxique, les contraintes phonologiques, la structure prosodique.

I.1.1.L'information distributionnelle

Les mots appartiennent à une même catégorie syntaxique, dans la mesure où deux mots de ce groupe sont interchangeables. Leurs caractéristiques sémantiques et phonologiques seront différentes, mais pas leurs propriétés syntaxiques. Ainsi, le mot *justice* est un nom, car il peut être remplacé par les mots *chien* ou *métallurgie*, sans que le contexte grammatical soit faux. En Anglais, les noms sont souvent précédés par '*the'*, et d'autres articles. En outre, chaque catégorie est définie à partir des autres catégories.

Maratsos et Chalkley (1980) ont imaginé un réseau de corrélation entre les mots, les morphèmes et le sens (Finch et Chater, 1992 ; Mintz, Newport et Bever, 1995). Par exemple, Mintz et coll. (1995) ont montré que la similarité entre les contextes peut être utilisée pour regrouper les mots, et que les groupes ainsi formés correspondent aux classes grammaticales. Une fenêtre utilisant les deux mots situés juste avant et après le mot à classer permet d'identifier les noms et les verbes. Ces travaux ont été complétés par Redington, Chater et Pinson (1998), où ils montrent que cette méthode est efficace pour les noms et les verbes, mais moins performante pour les mots de fonction.

I.1.2.L'amorçage sémantique

Le contexte sémantique (i.e. sens des mots) peut permettre de trouver les catégories grammaticales des mots (Pinker, 1984 ; Bates et MacWhinney, 1989). Ainsi les noms sont reconnus de part leur sens concret, et les verbes comme les mots désignant des actions (Braine, 1976 ; Schlesinger, 1971). Des expériences ont montré que les enfants utilisaient cette distinction (Cassidy et Kelly, 1991). Certains chercheurs ont envisagé que les premières étapes du développement grammatical contiennent uniquement des mots lexicaux (Guilfoyleet Noonan, 1992 ; Platzack, 1990 ; Radford, 1990). Ce n'est qu'à partir de 2 ans, que les enfants font usage des catégories fonctionnelles (Gerken, 1994).

I.1.3.Les contraintes phonologiques

La linguistique considère traditionnellement que la phonologie et les catégories grammaticales constituent des structures totalement indépendantes (Saussure, 1916/49 ; Chomsky, 1975 ; Hockett, 1966). Si effectivement une catégorie grammaticale ne peut être réduite à un ensemble de propriétés phonologiques, ces classes ont des propriétés phonologiques partiellement prédictibles (Kelly, 1992 ; Cassidy et Kelly, 1991).

La morphologie et la syntaxe formelle permettent également d'effectuer cette classification. Ainsi les verbes se terminent souvent par le suffixe '-ed', ou '-s'. Les verbes anglais contiennent également moins de syllabes que les noms. Deux expériences mettent en évidence que les enfants et les adultes sont sensibles à cette différence (Cassidy et Kelly, 1991). La phonologie peut être particulièrement utile pour des phrases courtes ou constituées d'un mot unique, qui disposent de moins d'indices morphologiques et syntaxiques. Cette situation n'est pas rare dans le cas du discours avec l'enfant (15 % dans le corpus utilisé par Cassidy et Kelly, 1991).

I.1.4.L'hypothèse d'amorçage prosodique

Pour apprendre une langue, il est possible de se fier aux indices acoustiques et prosodiques afin de résoudre un certain nombre de tâches comme la segmentation de la parole, ou la création des catégories syntaxiques (Gleitman, Gleitman, Landau et Wanner, 1988 ;Gerken, Jusczyk et Mandel, 1994 ; Jusczyk et coll., 1992 ; Kemler Nelson et coll., 1989 ; Morgan, 1986). La structure prosodique ne peut donc se définir uniquement à partir de la syntaxe. Ainsi, la prosodie fournit uniquement des indications, des contraintes qui permettent de guider l'analyse de la syntaxe. L'utilisation des propriétés segmentales abstraites du langage provient de l'apprentissage des premiers mots. Les enfants peuvent distinguer un certain nombre de mots en s'appuyant sur les propriétés acoustiques globales sans l'aide d'une représentation phonologique (Gerken, 1994).

L'hypothèse d'amorçage prosodique rejoint celle s'appuyant sur des contraintes phonologiques, mais ne nécessite a priori aucune connaissance linguistique, puisque les données prosodiques sont disponibles dans les caractéristiques acoustiques du signal. En outre, il va de soi que les quatre hypothèses proposées se complètent pour permettre l'acquisition de la syntaxe. Cependant, notre étude se cantonnera à l'utilisation de la prosodie pour identifier deux catégories syntaxiques de base : les mots de fonction et les mots de contenu.

I.2.Définition des mots de fonction et de contenu

La distinction entre les mots grammaticaux et lexicaux est supposée être universelle (Abney, 1987). En outre, ces deux types remplissent des fonctions opposées au sein de l'organisation syntaxique : les mots de fonction établissent la structure syntaxique tandis que les mots de contenu définissent les composants sémantiques insérés dans cette structure syntaxique.

I.2.1.Les mots de contenu

Les mots de contenu possèdent un réel contenu sémantique : les noms, les verbes, les adjectifs et les adverbes (Shi, Morgan et Allopenna, 1998). Il existe dans chaque langue un très grand nombre de mots de contenu (plusieurs dizaines de milliers), et ce nombre est en principe illimité : de nouveaux mots apparaissent tous les jours en fonction des nouveaux concepts que nous avons besoin d'exprimer. Pour cette raison ils sont aussi dénommés mots de classe ouverte, ou mots lexicaux. La plupart des mots de contenu apparaissent rarement.

I.2.2.Les mots de fonction

Tous les autres mots sont dénommés mots de fonction ou mots grammaticaux. Il s'agit des pronoms, déterminants, prépositions, verbes auxiliaires, compléments, conjonctions, déterminant, (quelques) prépositions et mots de question. On y inclut aussi les morphèmes de fonction, tels ceux qui marquent le nombre, le genre et autres propriétés grammaticales, et qui sont généralement attachés aux mots de contenu, particulièrement dans le cas des langues dites agglutinantes. Cependant, cette liste peut varier d'une langue à l'autre (Shi et coll., 1998).

Ces mots ont un contenu sémantique faible ou indéfinissable, et sont en nombre très limité et fixe. C'est pourquoi, ils sont appelés aussi les mots de classe fermée. Les mots de fonction servent à relier les mots de contenu et à véhiculer la structure syntaxique de la phrase. En conséquence, les mots de fonction apparaissent très souvent. Cette fréquence élevée d'utilisation leur confère une tendance à être minimum d'un point de vue acoustique et/ou phonologique (Shi et coll., 1998).

Maintenant que l'objectif de cette section est clairement présenté, un point sur les rapports entre prosodie et syntaxe sera examiné (section II), avant de décrire le matériel et les méthodes (section III) qui ont été retenus pour la partie expérimentale (section IV). Nous discuterons alors des implications de ces travaux (section V).

II.Le contexte de la catégorisation lexicale

Jusczyk et Kemler Nelson (1996) ont proposé que les enfants en bas âge puissent utiliser immédiatement leur sensibilité aux marqueurs prosodiques afin d'organiser les entrées auditives. Plusieurs investigations ont été réalisées pour confirmer cette hypothèse dans le cadre de l'acquisition de la syntaxe :

1. des descriptions des indices prosodiques, distinguant des catégories lexicales comme les mots de fonction et de contenu, ou les noms et les verbes ;
2. des études de la sensibilité des bébés pour les indices prosodiques ;
3. des systèmes automatiques d'identification des catégories lexicales, à partir d'un ensemble d'indices, dont des indices prosodiques.

II.1.Distinction phonologique et prosodique de catégories lexicales

II.1.1.Les mots de fonction et de contenu

Les mots de fonction et de contenu possèdent généralement des propriétés phonologiques assez différentes (Selkirk, 1984 ; Nespor et Vogel, 1986). Swanson, Leonard et Gandour (1992) ont suggéré que les variations apparaissent surtout sur les mots de contenu, au détriment des mots de fonction. De manière plus générale, Selkirk (1996) propose deux propriétés universelles des mots de fonction :

1. Contrairement aux mots de contenu, les mots de fonction n'engendrent pas systématiquement leurs propres mots prosodiques.
2. Les mots de fonction ont une prosodie plus variable que les mots de contenu à la fois entre les langues et au sein d'une langue.

Cutler (1993) suggère que les mots de fonction aient des voyelles réduites dans leur première syllabe, alors qu'elles ne le sont pas dans les mots de contenu. Certaines propriétés des mots de contenu semblent quasiment universelles. En Chinois Mandarin, en Turc et en Anglais les mots de fonction tendent à être minimaux au sens de la perception, tout en respectant les règles phonologiques de leurs langues. De ce fait, les mots de fonction ont une structure syllabique plus simple, tandis que les mots de contenu ont le profil inverse. Cette minimalisation s'applique également au niveau prosodique. Ainsi, ils ont des syllabes avec des durées réduites et une intensité relative plus faible (Shi et coll., 1998).

En Français, « si un mot de fonction suit un mot de contenu, alors ce mot de contenu est à la fin d'un groupe intonatif et reçoit un accent final » (Malfrère, Dutoit et Mertens, 1998). En Anglais, les mots de fonction monosyllabiques ne sont pas souvent accentués (Gleitman et Wanner, 1982) alors que les mots de contenu sont marqués par un accent principal et ne sont jamais réduits (Hirst et Di Cristo, 1998). En Serbo-croate, les mots de contenu comportent toujours un ton élevé sur l'une de leurs syllabes, mais pas les mots de fonction. En Japonais (Tokyo), les mots de fonction peuvent perdre leur ton élevé dans certaines circonstances, les mots de contenu jamais (Venditti, Jun et Beckman, 1996 ; Selkirk, 1996 ; voir Peters et Strömqvist, 1996 ; Morgan et coll., 1996 ; Hung et Peters, 1997 pour des exemples en Suédois, Chinois Mandarin, Taïwanais...).

Pour l'Anglais, la production des accents la plus simple est de supposer que chaque mot de contenu porte un accent. Cependant, la position de l'accent est influencée par des facteurs sémantique, syntaxique et pragmatique. Ainsi, tous les mots de contenu ne sont pas accentués en Anglais, particulièrement lorsqu'il s'agit de textes plus longs. La prédiction de l'accent dans les groupes de noms communs en Anglais est particulièrement difficile, parce qu'elle dépend du rôle sémantique. De même, l'ordre sériel des élèments dans la phrase détermine quels mots sont accentués. Le statut du mot dans le discours participe également à la prédiction des accents. Ainsi, un mot nouvellement employé dans la conversation aura tendance à être accentué, mais ses occurrences

futures ne seront plus accentuées (Hirschberg, 1993).

Le discours adressé aux enfants peut contenir des indices acoustiques facilitant l'acquisition du langage. Par exemple, l'espace vocalique est augmenté dans le cas du discours vers l'enfant. D'un point de vue spectral, les fréquences F1 et F2 sont analysées pour les voyelles /i/ /a/, /u/. Le triangle vocalique s'élargit pour les mots de contenu, lorsque le locuteur s'adresse à un enfant. Pour les mots de fonction le contraire est vérifié. Le triangle est plus étendu pour les mots de fonction lorsque le discours est adressé aux adultes (Van de Weijer, 2001). Mais tous les locuteurs ne présentent pas l'expansion de l'espace vocalique (la baby-sitter et la mère étendent le triangle vocalique).

II.1.2.Les noms et les verbes

Les différences observées entre les mots de fonction et les mots de contenu se retrouvent pour des catégories syntaxiques plus fines comme les noms et les verbes. Effectivement, Kelly (1992) a remarqué, pour l'Anglais, que les noms dissyllabiques ont tendance à être accentués sur la première syllabe, tandis que les verbes dissyllabiques sont accentués sur la seconde. Ces différentes propriétés semblent être utilisées par des adultes, ainsi que par des enfants de quatre ans, pour distinguer ces deux catégories grammaticales.

Francis et Kucera (1982) ont montré que 94 % des noms ont un accent trochaïque, alors que 69 % des verbes ont un accent ïambique. En outre, 85 % des mots avec un rythme ïambique sont des verbes, et 90 % des mots avec un rythme trochaïque sont des noms. Les corrélats phonologiques des catégories noms et verbes contiennent : la durée, la réalisation acoustique des phonèmes, et le nombre de phonèmes.

Kelly (1996) indique que les noms ont tendance à avoir plus de syllabes que les verbes en Anglais. Pour un corpus de discours adressé à l'enfant, la probabilité d'avoir un nom monosyllabique est seulement de 38 %, pour les mots de deux et trois syllabes la probabilité dvient respectivement 76 % et 94 %. Tous les mots ayant quatre syllabes sont des noms. En outre, les noms et les verbes de même longueur syllabique n'ont pas la même durée (Cassidy et Kelly, 1991). Les noms ont des voyelles plus graves, plus de consonnes nasales, et plus de phonèmes (pour un même nombre de syllabe).

Les deux sous-sections précédentes ont illustré un certain nombre de différences phonologiques et prosodiques entre différentes catégories lexicales. Les enfants et les nourrissons peuvent-ils révéler des comportements, montrant qu'ils réagissent à ces indices et les exploitent pour la syntaxe ?

II.2.Sensibilité aux structures prosodiques pour la catégorisation lexicale

Les mots de fonction et de contenu se distinguent par leurs propriétés phonologiques, prosodiques et acoustiques. Les sujets américains adultes peuvent exploiter des indices phonologiques, pour identifier des mots inconnus comme des verbes ou des noms (Kelly, 1996). Le chapitre 2 a présenté la sensibilité des êtres humains pour certains indices prosodiques, qu'en est-il pour la syntaxe chez les enfants et les nouveau-nés ?

II.2.1.Chez l'enfant

Les premières phrases des enfants ont un style télégraphique, parce qu'ils n'utilisent pas les mots de fonction, et mettent en avant les syllabes fortes des mots de contenu. Les omissions augmentent au fur et à mesure que la phrase à produire est longue (Gerken, 1994). Ce constat laisse penser que le langage des jeunes enfants s'appuie sur les mots de contenu. Sous cette hypothèse les enfants se fient aux mots de contenu, et considèrent les mots de fonction comme un bruit.

Ce phénomène peut s'expliquer de deux manières :

1. Les mots de contenu sont appris en priorité de part leur nature concrète.
2. Les enfants prêtent attention aux mots et syllabes accentués.

En Quiché maya, les morphèmes de fonction reçoivent un accent plus marqué que les mots de contenu. Les enfants ont alors tendance à omettre les mots de contenu et à conserver les mots de fonction plus souvent que dans les langues, qui accentuent les mots de contenu au détriment des mots de fonction (Pye, 1983 cité dans Gerken, 1994). Les enfants dès l'âge de 2 ans ont une représentation de quelques morphèmes de fonction spécifiques et du contexte dans lesquels ils apparaissent. Effectivement, les enfants anglais omettent plus fréquemment les mots de fonction que des syllabes non accentuées n'ayant aucun sens. En outre, les enfants perçoivent les mots de fonction sous deux formes phonologiques possibles : (+fricative) schwa (+fricative) ou schwa (+ nasal). Les enfants sont sensibles aux motifs prosodiques de leurs langues, aussi bien au niveau des mots, que des phrases (Gerken et McIntosh, 1993).

Pendant la période relative à la production de un à deux mots, les mots de contenu reçoivent une articulation plus précise, dans le discours des parents. Lorsque l'enfant prononce plus de deux mots, les éléments grammaticaux ont alors une énonciation plus précise (Ratner, 1984). Le discours des parents s'adapte à l'enfant en fonction de ses compétences syntaxiques. Les enfants sont capables d'exploiter des indices divers de la parole pour faire une distinction entre ces différentes catégories lexicales. Nous allons maintenant considérer la capacité des nouveau-nés pour exploiter de tels indices pour obtenir les premières bases syntaxiques.

II.2.2.Chez les nouveau-nés

Les nouveau-nés ont des capacités remarquables pour les détails acoustiques et phonétiques du langage (Jusczyk, 1997). Dès 6-8 mois, ils savent utiliser des informations probabilistes, et des motifs de cooccurrence, pour trouver les structures et les règles. Ils discriminent les syllabes, d'après les catégories phonétiques, la forme et le nombre de syllabes et l'intonation.

Les nouveau-nés (âgés de seulement trois jours) peuvent employer des combinaisons probabilistes d'information acoustique et phonologique pour séparer perceptuellement des mots anglais dans des catégories de fonction/contenu (Shi, Werker et Morgan, 1999). Les nouveau-nés ont été testés avec un paradigme d'habituation et une

procédure de succion de haute amplitude. Deux listes de mots isolés leur sont présentées. Elles contiennent soit des mots de fonction, soit des mots de contenu [92] . Les nouveau-nés distinguent plus facilement des listes de mots issus de catégorie grammaticales différentes. Etant donné que dans la première expérience, seuls les mots grammaticaux ne pouvaient posséder qu'une syllabe, une seconde expérience a fait appel à des listes de mots lexicaux ne contenant qu'une seule syllabe. Effectivement, Bijeljac-Babic, Bertoncini et Mehler (1993) avaient déjà démontré que les nourrissons savent faire la différence entre des mots ayant un nombre différent de syllabes. Sous ces nouvelles conditions, les auteurs retrouvent la même distinction. Ultérieurement, de telles capacités peuvent aider les enfants pour détecter et représenter des classes des mots sur la base de ces indices perceptuels. Dans cette expérience, la discrimination s'opère à l'écoute de mots déjà segmentés. Des expériences récentes ont testé les capacités des nourrissons pour la segmentation des mots de fonction comme « a » et « the » dans la parole continue, capacités qui n'apparaissent qu'à partir de 10 mois et demi (7 mois et demi pour des noms ; Shady, Jusczyk et Gerken, 1998).

Shafer, Gerken, Shucard, et Shucard (1992) et Shafer, Shucard, Shucard et Gerken (1998) ont par ailleurs montré que des enfants de 11 mois avaient des potentiels évoqués différents lors de l'écoute de phrases où les mots de fonction avaient été remplacés par des syllabes arbitraires, par rapport à l'écoute de phrases normales. Un résultat similaire a également été obtenu par Shady et coll. (1998) avec des enfants de 10 mois et demi en utilisant la technique d'orientation préférentielle de la tête. Ces auteurs ont de plus répliqué ce résultat en utilisant de faux mots de fonction qui avaient des propriétés phonologiques similaires aux mots d'origine. Ce résultat suggère qu'à 10 mois et demi, les enfants connaissent déjà un certain nombre de mots de fonction, et ne se fient plus seulement à leurs propriétés phonologiques générales.

Si les nourrissons sont capables d'exploiter un certain nombre d'indices pour différencier des catégories syntaxiques, il doit être possible de concevoir des systèmes automatiques exploitant également ces indices pour trouver certaines catégories lexicales.

II.3.Etat de l'art de l'identification de catégories lexicales

Trois systèmes différents proposés pour l'identification des catégories lexicales vont être examinés :

1. Le premier a pour architecture une carte auto-organisatrice et distingue les catégories qui nous préoccupent (i.e. fonction et contenu) ;
2. Le second présente un apprentissage statistique supervisé d'exemple de mots de différentes catégories syntaxiques ;
3. Le dernier s'appuie sur un réseau récurrent simple (SRN).

[92] Les catégories lexicales de ces mots ont été retrouvées préalablement par des cartes auto-organisatrice de Kohonen (*cf.* II.3.1 ; Shi et coll., 1998).

II.3.1.A partir de carte auto-organisatrice (Shi et coll., 1998)

Shi et coll. (1998) ont étudié si divers indices présyntactiques sont suffisants pour guider l'attribution des mots aux catégories grammaticales rudimentaires. Leur recherche sur l'Anglais, le Chinois Mandarin et le Turc prouve que des ***" ensembles d'indices distributionnels, phonologiques, et acoustiques distinguant les articles lexicaux et fonctionnels sont disponibles dans le discours dirigé à l'enfant au travers de langues de typologie distincte telles que le Mandarin et le Turc "*** (Shi et coll., 1998). Leur étude est effectuée à partir d'un corpus de discours adressé à l'enfant [93] . 5 % des mots sont tirés au hasard puis transcrits pour faire partie de l'analyse [94] . Chacun de ces mots est représenté par un ensemble d'indices (certains indices sont spécifiques de la langue étudiée) :

Mesures distributionnelles : 1.

- Fréquence du type (indice lexical) ;
- Position dans la phrase (début, milieu, ou fin).

Mesures phonologiques : 1.

- Nombre de syllabes [95] ;
- Structure syllabique [96] ;
- Présence d'une nasale en fin de syllabe (Coda) ;
- Duplication d'une syllabe (Mandarin : plus souvent sur les mots de contenu) ;
- Ton appuyé (Mandarin uniquement) ;
- Harmonie vocalique (Turc uniquement).

Mesures acoustiques (calculées pour une syllabe, puis moyenne effectuée sur le mot)1.

- Durée de la syllabe ;
- Amplitude relative [97] ;
- Variation de F0 [98] (calculée en demi-tons, et normalisée par la durée).

La durée des syllabes a été examinée seule pour le Mandarin. Elle permet, moyennant un apprentissage supervisé, une identification de 71 % et 90 % pour chacune des deux mères. La durée séparant le mieux les mots de fonction et de contenu est de 135 ms pour la première mère, 125 ms pour la seconde. Trois apprentissages non supervisés ont été

93 Deux mères avec deux enfants âgés de 11 et 20 mois.

94 En Anglais, on dénombre 67 mots de contenu et 31 de fonction pour la première mère et 49 mots de contenu et 28 mots de fonction pour la seconde mère.

95 Mandarin : 29 % de l'extrait des mots lexicaux sont multisyllabiques, 49 % dans la totalité ; Turc : 50 % des mots de contenu sont multisyllabiques. 91 % des mots de fonction ne contiennent qu'une seule syllabe.

96 Mandarin : 65 % des mots de contenu ont des diphtongues. 19 % pour les mots de fonction.

97 Ratio énergie RMS de la syllabe courante par la syllabe la plus forte de la phrase.

98 Aucune différence significative n'est observée entre les deux catégories, mais les mots de contenu varient plus que les mots de fonction.

appliqués sur la durée des syllabes de la seconde mère (Cluster : 66 %, division suivant lamoyenne de tous les mots : 81 %, médiane : 84 %),et de la première mère (respectivement : 56 %, 60 % et 58 %). L'indice de durée des syllabes est l'indice qui obtient le pourcentage d'identification correcte le plus élevé.

L'utilisation de carte auto-organisatrice de Kohonen (1982) permet d'identifier les mots de fonction des mots de contenu à partir de l'ensemble de ces indices exceptés les variations de F0. Chacune des entrées est normalisée, si bien que toutes ont le même poids pour le réseau. Les unités du réseau sont étiquetées en fonction de leur réponse, pendant l'apprentissage [99] . Pour le Mandarin, les performances sont les suivantes : Mère 1 : 93 %, Mère 2 : 88 %. En outre, les deux catégories ont le même degré de reconnaissance. Pour le Turc, les performances atteignent 86 % et 84 % pour chaque mère.

Les enfants pourraient classer des mots segmentés ou des morphèmes dans des super-catégories, avant de savoir le sens de ces mots et d'avoir une représentation de l'analyse distributionnelle des mots présents dans le signal de parole. Certaines propriétés semblent universelles (durée des voyelles ou des syllabes plus courtes, moins de syllabes pour les mots de fonction).

II.3.2.Apprentissage à partir d'exemple (Durieux et Gillis, 2000)

Durieux et Gillis (2000) ont proposé un système artificiel pour assigner des classes grammaticales, en utilisant diverses informations phonologiques et prosodiques. Leur système apprend un certain nombre d'exemples, et détermine pour un nouvel item, lequel des exemples appris en est le plus proche. Cet apprentissage dénommé « Lazy Learning » est utilisé dans une procédure de validation croisé (ou « leaving one out »).

Leur premier ensemble testé contient 212 noms, 215 verbes, et 16 formes ambiguës, (tous dissyllabiques et homographes). La position de l'accent (noté de façon discrète) permet d'identifier les noms et les verbes avec un score de 82.6 %. Lorsque cet ensemble est étendu à 5000 mots non homophones et contenant jusqu'à quatre syllabes, les performances chutent, en particulier pour l'identification des verbes (< 40 %). D'autres différences pourraient distinguer les noms des verbes, ainsi le premier « s » de « uses » n'est pas voisé, ou la première voyelle de « cashier » est réduite à un schwa dans le cas d'un verbe.

Une autre expérience est proposée à partir de l'ensemble des indices phonologiques mis à jour par Kelly (1996). Les performances atteignent 78.2 % pour 5000 mots tirés au hasard. Un indice isolant les noms des verbes en Anglais peut être moins prédictif dans une autre langue (de 66 % en Anglais à 58 % en Néerlandais pour l'accent).

Leur dernière expérience porte sur l'identification de quatre catégories (noms, verbes, adjectifs et adverbes) pour deux langues. Les scores sont 66.7 % et 71 %, deux performances supérieures à un tirage aléatoire. L'utilisation combinée des indices

[99] Trois choix possibles en fonction des unités du voisinage. Si l'une des unités du voisinage donne une réponse différente, la réponse est considérée comme confuse et n'est pas prise en compte. Si toutes les unités ne sont pas étiquetées, le mot testé n'est pas classé.

phonologiques et prosodiques permet d'obtenir les meilleurs résultats pour l'identification de plusieurs classes grammaticales de mots de contenu, pour l'Anglais ainsi que le Néerlandais.

II.3.3.Réseau Récurrent Simple (SRN, Reali, Christiansen et Monaghan, 2003)

Christiansen et Dale (2001) ont montré les capacités du réseau SRN, proposé par Elman (1990), pour l'apprentissage des catégories lexicales sur un mini-langage abstrait. Le réseau SRN apprend la catégorie lexicale suivant l'entrée courante. Les informations passées sont encodées lors de l'apprentissage.

Ces résultats ont été étendus par Reali et coll. (2003) sur un corpus de parole adressé à l'enfant (Berstein-Ratner, 1984). Le vocabulaire est très réduit, et contient seulement 15 catégories lexicales. Il ne s'agit pas d'assigner un mot donné à une catégorie lexicale, mais plutôt d'apprendre l'organisation des catégories syntaxiques. Cette simulation est éloignée de la réalité, car un enfant ne peut pas disposer des catégories lexicales, au mieux il peut les déduire à partir de ces connaissances et du signal de parole.

Le SRN est donc appliqué au traitement de 16 indices phonologiques, mis à jour par Monaghan, Chater et Christiansen (2003). Dans une première étape, les noms et les verbes sont identifiés à l'aide d'une analyse discriminante. Le nombre de syllabes est l'indice donnant les meilleures performances (57.4 %). En utilisant tous les indices, les verbes et les noms sont identifiés avec un taux de 76 %. Dans une seconde étape, le réseau SRN a été entraîné sur les représentations phonétiques des mots. L'analyse discriminante des unités cachées du réseau permet un taux d'identification des noms et des verbes de 86 %. Les auteurs en concluent en outre que le réseau a su utiliser les informations distributionnelles en plus des informations phonétiques pour effectuer cette disctinction.

Dans toutes les études présentées ici, un certain nombre d'indices spécifiques ont été extraits à partir du discours avec l'aide supplémentaire d'un expert humain. Les enfants sont censés avoir une représentation des mots ou des morphèmes individuels. Ici nous étudierons si un système s'appuyant sur des données neuro-réalistes (TRN) peut automatiquement extraire de tels indices à partir du signal de parole lui-même. La section suivante présente les méthodes et le matériel employés pour répondre à ces contraintes.

III.Matériel et méthodes

Cette partie s'articulera autour de trois points : 1. les corpora employés (MULTEXT et LSCP), 2. l'obtention de valeurs de la F0, 3. Les méthodes de traitement de ces valeurs.

III.1.Corpora

Chacun des deux corpora est segmenté manuellement en mots. Cette étape est

nécessaire, et à l'heure actuelle aucune étude n'a entrepris d'identifier la catégorie lexicale d'un mot, sans savoir auparavant à quelle partie du signal il correspond. En outre, nous avons pris en compte deux segmentations. La première est basée sur les mots d'une même catégorie lexicale. Ainsi, deux mots de contenu se suivant ne font qu'un seul groupe. Dans ce cas les groupes de mots de fonction et de contenu se succèdent alternativement. La seconde emploie les mots eux-même, donc deux mots de contenu peuvent se succéder.

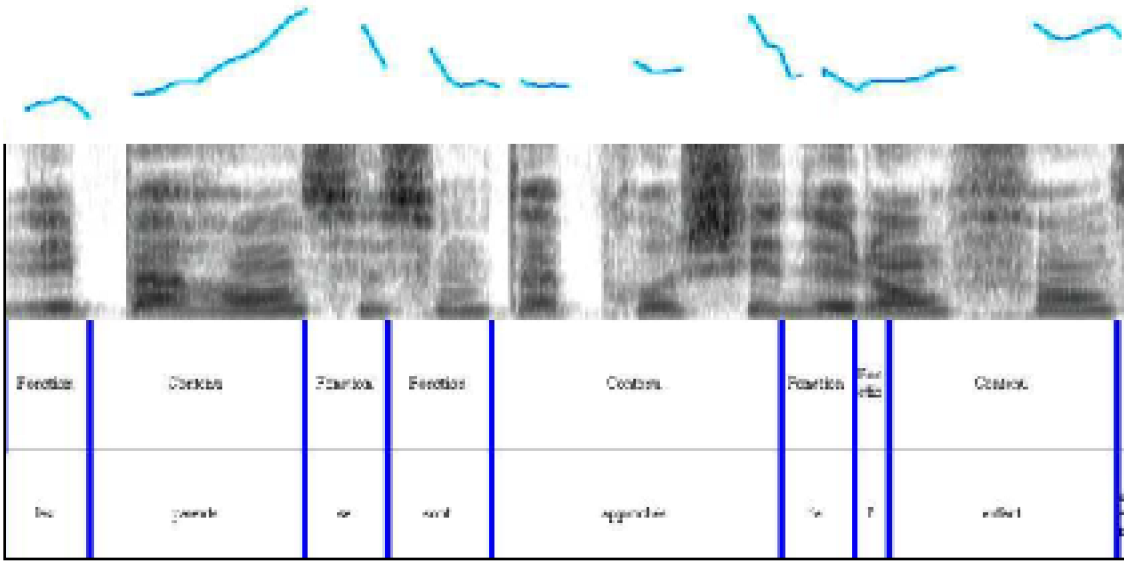


Figure 5.1 Exemple de segmentation manuelle d'une phrase, réalisée par C. Dodane (Tout en haut est figurée la F0, et en dessous un spectrogramme).

III.1.1.LSCP

Pour le Français, ce corpus contient 54 phrases françaises lues par un unique locuteur. Une segmentation fournit les positions dans le signal de parole des groupes de mots de contenu et de fonction (~ 200 pour chaque catégorie). Ces phrases faisaient partie du corpus multilingue établi par le laboratoire LSCP (Ramus et coll., 1999). Cette segmentation a été étendue aux mots eux-même par Christelle DODANE, ainsi que pour un locuteur Anglais. Les phrases anglaises sont des traductions approximatives des phrases françaises. La partie française du corpus fait l'objet des investigations de la première section (IV.1), sauf indication contraire.

III.1.2.MULTEXT

Les corpora Français et Anglais ont été extraits du corpus multilingue MULTEXT développé pour l'étude de la prosodie (Campione et Veronis, 1998). Des histoires ont été lues par 20 lecteurs différents (5 hommes et 5 femmes par langue) pour un total de 8236 mots pour l'Anglais, et 6945 mots pour le Français.

Les catégories syntaxiques des mots présents dans le corpus ont été déterminées à l'aide des logiciels more et post (pour la désambiguïsation) de la boîte à outils CLAN

utilisée dans les spécifications CHILDES (www.childes.com).

III.2.Représentation des données

Nous examinerons deux types de valeurs de F0, obtenues avec deux logiciels de traitement du signal différents. Nous ferons intervenir uniquement deux catégories phonétiques : consonnes, voyelles. Chacune de ces catégories est encodée par un seul neurone de la couche d'entrée (section IV.2.1).

Pour les premières expériences avec le corpus LSCP, la fréquence fondamentale est extraite par intervalles de 10 ms, en utilisant le logiciel BLISS, de John Mertus. Ces valeurs ont été transmises par le laboratoire LSCP. Les valeurs de la fréquence fondamentale (F0 : données brutes) ont également été obtenues par l'autocorrélation du signal chaque 5 ms, (logiciel de PRAAT ; Boersma 1993). Ces valeurs de F0 ont été transmises au TRN, par l'intermédiaire d'une courbe de Gauss.

Les valeurs brutes de la F0 subissent un traitement de façon à obtenir une représentation proche de l'impression laissée par la perception. L'algorithme MOMEL (Hirst et Espesser, 1993) a été employé pour obtenir une représentation perceptuelle acceptable de l'intonation à partir des valeurs brutes de F0. L'application d'une courbe continue lisse (basée sur des fonctions splines quadratique) reflète le contour intonatif de la parole. Une description plus détaillée de cet algorithme a fait l'objet du chapitre 2 (section III.2.2) consacré à la prosodie. Les commandes Interpolate et Smooth du logiciel PRAAT seront aussi employées pour obtenir une représentation continue et lissée des valeurs de F0.

Nous utiliserons également un spectrogramme (basé sur une échelle de perception Mel ou linéaire) pour représenter la partie prosodique dédiée à la fréquence fondamentale. Ainsi, la première couche d'entrée du réseau est constituée par une représentation spatio-temporelle du signal. Nous envisageons d'étudier trois représentations des fréquences inférieures à 400 Hz :

1. Un spectrogramme avec une fenêtre d'analyse de 80 ms en ne tenant compte que des valeurs inférieures à 400Hz. La résolution fréquentielle est fixée à 2.75 Hz, et conduit à 143 neurones d'entrées. Chaque valeur est multipliée par 500. Nous tiendrons compte d'un spectrogramme avec les fréquences inférieures à 5000 Hz et résolution de 20 Hz conduisant à 256 neurones d'entrées.
2. un cochléogramme : La résolution fréquentielle est fixée à 10 Barks, la taille de la fenêtre d'analyse est de 30 ms, la fenêtre de masquage rétrograde est de 30 ms également. Ces valeurs sont les valeurs par défaut du logiciel PRAAT.
3. Un spectrogramme dont les valeurs des filtres sont alignés sur une échelle de perception Mel (algorithme Melfilter de PRAAT) inférieures à 500 mels avec une fenêtre d'analyse de 60 ms. La résolution est de 12.5 mels ce qui conduit à 40 neurones d'entrées.

III.3.Méthodes de traitement

Pour identifier les catégories, nous avons tenu compte de méthodes « classiques » pour mettre en relief les indices potentiels pour l'identification des mots de fonction et de contenu, avant d'effectuer cette catégorisation lexicale à partir de la fréquence fondamentale traitée par le TRN.

III.3.1.Analyse des données

Parmi les premières méthodes testées (*cf.* section IV.1) nous avons des méthodes issues soit des statistiques (prototype moyen et analyse discriminante utilisation du logiciel SPSS), soit du connexionnisme (réseaux probabilistes, cartes auto-organisatrice de Kohonen, 1982).

Le réseau probabiliste (désigné sous l'abréviation pnn dans la boîte à outils réseau de Matlab) apprend les exemples de la base d'apprentissage. Il faut ensuite déterminer le rayon qui permet d'obtenir les meilleures performances en validation. Pour cet apprentissage, nous n'avons pas testé systématiquement avec un corpus de développement supplémentaire, pour pratiquer la validation en aveugle. Dans cas, la performance annoncée indique un seuil maximal d'identification.

Des cartes auto-organisatrice de Kohonen (abrégées par SOM, Boîte à outils réseaux de MATLAB) seront également étudiées. Ce dernier apprentissage est non supervisé et différents paramètres (comme le nombre de cycles d'apprentissage, l'architecture de la carte) seront indiqués pour chaque utilisation. Dans chacun de ces cas, nous aurons recours à deux sous-ensembles distincts d'apprentissage et de validation, couvrant chacun la moitié des valeurs disponibles dans le corpus. Durant l'apprentissage, le neurone qui répond à un mot testé se voit assigner la catégorie de ce mot. Dans le cas où un neurone répond aussi à bien des mots de l'une ou l'autre des catégories, la catégorie à laquelle il a répondu le plus souvent lui est assignée. Nous donnons alors les performances d'identification de ce réseau pour le corpus de validation.

III.3.2.Le réseau récurrent temporel (TRN)

Le fonctionnement du réseau récurrent temporel a été présenté dans le premier chapitre (*cf.* Chapitre Un, section 3.3). Le réseau a été testé avec divers types d'entrées : soit uniquement rythmiques avec une représentation simple des catégories phonémiques (un neurone d'entrée représente une catégorie), soit intonatives à partir de la représentation de la fréquence fondamentale employée dans les deux chapitres précédents. Contrairement aux chapitres précédents, l'état du réseau ne sera pas relevé à la fin d'une phrase, mais à la fin de chaque mot. Nous tiendrons compte des deux couches cachées (State et $State_D$) pour représenter un mot. Ainsi, chaque mot sera encodé par une représentation qui décrira la représentation intonative du mot courant.

Nous présenterons également les moyennes des performances obtenues, sur chaque moitié du corpus LSCP, en phase de validation. Effectivement, le petit nombre de données disponible pour ce corpus influence les résultats.

IV.Expérimentation

Cette partie décrit les diverses expériences pour la discrimination des mots de fonction et de contenu. Nous avons dans un premier temps considéré des méthodes classiques pour identifier ces mots à partir de l'information acoustique, afin d'établir les indices susceptibles d'accomplir cette identification.

Nous nous intéresserons ensuite à appliquer le TRN à cette tâche de classification. Ainsi nous validerons le fait que le réseau TRN puisse traiter des informations temporelles définies dans une échelle temporelle courte, en respectant notre contrainte temporelle. La catégorisation lexicale sera donc effectuée à partir des données acoustiques non étiquetées, hormis les mots.

IV.1.Détermination d'indices pour la catégorisation lexicale

Ce paragraphe résume les techniques utilisées pour identifier les catégories lexicales (fonction et contenu) à partir des mots considérés comme des segments isolés. Dans ce cas, les frontières de début et de fin encadrant le mot sont prises en compte. Le réseau TRN n'est pas utilisé. Nous étudierons les performances obtenues par des méthodes classiques soit en statistique avec l'évaluation obtenue en calculant la distance entre un item et un prototype moyen, l'analyse discriminante (utilisation du logiciel SPSS) ; soit en connexionisme avec des réseaux probabilistes.

Les données sont représentées au cours du temps avec un échantillonnage de 10 ms. Dans un premier temps, seules les valeurs de la fréquence fondamentales sont employées. A partir de ces données nous extrayons un certain nombre de statistiques pour caractériser les mots, par exemple la position du maximum de F0, la valeur moyenne de la F0. Nous obtenons une matrice où sont indiqués la catégorie des mots, leurs durées et les valeurs obtenues pour les statistiques. Sauf mention contraire, le corpus considéré sera le corpus LSCP Français.

Tout d'abord, les informations de durée et les valeurs de F0 seront considérées sur le domaine de la voyelle (IV.1.1). Ensuite, l'impact de la durée des mots (IV.1.2), et de leurs contours intonatifs seront distingués (IV.1.2). Les contours intonatifs seront par la suite réduits à la seule information de la présence d'un pic d'intonation (IV.1.4). Enfin, des prototypes du contour de F0, de l'intensité et des premiers formants seront pris en compte pour identifier en premier lieu les groupes de mots de même nature lexicale, puis les mots eux-même (IV.1.5).

IV.1.1.Indices vocaliques

Pour ces premiers investigations, seuls les groupes de mots consécutifs de même nature lexicale ont été étudiés. Nous retiendrons que deux types d'indices définis sur les voyelles (la durée et la fréquence fondamentale), avant d'étudier leur combinaison.

IV.1.1.1.Durée

Shi et coll. (1998) ont mis en avant l'importance de la durée des voyelles pour discriminer les mots de fonctions des mots de contenu. Morgan et coll. (1996) ont trouvé que la durée moyenne des voyelles contenues dans un mot permet un taux de 64 % d'identification pour l'Anglais.

Les durées que nous avons observées correspondent à des durées données dans d'autres études (Dodane et coll., en préparation). Nos premiers résultats montrent que les voyelles des mots de fonctions sont également prononcées plus rapidement par rapport aux mots de contenu, même si les données ne sont pas normalisées. Ces résultats rejoingnent l'hypothèse selon laquelle les mots de fonction sont minimaux (Morgan et coll., 1996). Le ratio des durées des consonnes par la durée des voyelles ne diffère pas significativement.

		Valeurs moyennes		Anova
Indices		Contenu	Fonction	F
Durée	Voyelles	0,079	0,053	0,000
	Consonnes	0,070	0,063	0,017
	totale	0,074	0,067	0,000
Ratio	Consonnes / Voyelles	0,761	0,869	0,204

Tableau 5.1 Valeurs moyennes de la durée des phonèmes, en ms.

La durée moyenne des voyelles sur un groupe de mots est prise comme prototype de chacune des deux classes. Le taux d'identification pour un corpus de validation (composé de la moitié des groupes de mots du corpus total) est de 71.5 %. Une carte de Kohonen à 2 dimensions (5 x 5) atteint une performance de 74.9 % après 150 cycles d'apprentissage.

Les performances sont supérieures à celles rapportées par Morgan et coll. (1996) pour l'Anglais (64 %). Toutefois, il ne s'agit pas des mots eux-même dans notre cas, et il est probable que la nature de la langue et des passages étudiés influence aussi les résultats.

L'algorithme de segmentation automatique en consonnes / voyelles développé par Pellegrino (1998) a été appliqué sur l'ensemble du corpus MULTEXT. Une analyse discriminante permet d'effectuer la tâche d'identification lexicale à partir de la durée des voyelles segmentées automatiquement (Anglais : 73.2 % ; Français 73.2 %). Ces observations sont valides si chaque groupe est considéré dans son entier. Est-il possible de ne plus utiliser qu'une seule voyelle ?

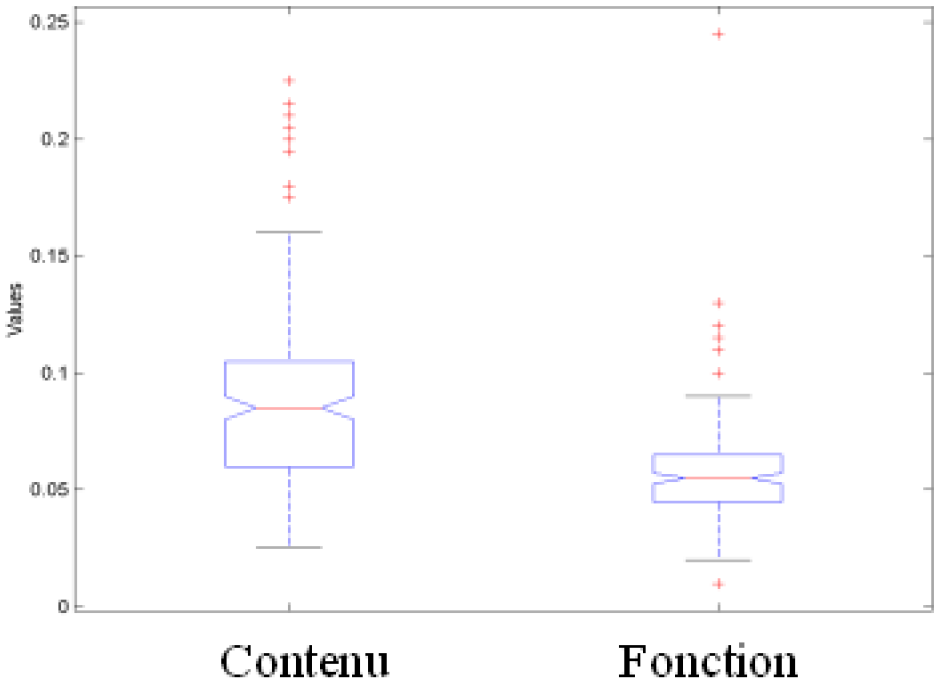


Figure 5.2 Durée moyenne de la dernière voyelle de chaque groupe.

Lorsque seule la dernière voyelle est prise en compte, la différence de durée entre les groupes de fonction et les groupes de contenu est significative (ANOVA : $p < 0.001$). Le taux de reconnaissance à partir de cette valeur moyenne est de 71 % pour la validation, 72 % avec un réseau probabiliste. L'utilisation d'une carte de Kohonen valide l'hypothèse d'un apprentissage non supervisé (10 neurones et 2500 cycles : 69,6 %).

Ces résultats montrent qu'il est possible de n'utiliser que la dernière frontière d'un groupe de mots sans que les performances d'identification soient trop atténuées.

Les observations précédentes indiquent que la dernière voyelle des groupes de mots semble un indice fiable pour l'identification en Français. Est-il possible de tenir compte de l'information apportée par la F0 ?

IV.1.1.2.Valeur moyenne de la fréquence fondamentale

La valeur moyenne de la F0 est calculée sur la dernière voyelle. La F0 est estimée par une technique d'autocorrélation toutes les 10 ms, avec PRAAT. Les groupes de mots de fonction prennent des valeurs de F0 plus restreintes, et leur fréquence moyenne est plus faible. Une analyse de variance révèle une différence significative entre les types de groupes ($P = 0.003$; cf. Figure 5.3). A l'aide d'un prototype moyenne, 63.2 % des groupes sont correctement identifiés. Un réseau probabiliste permet d'obtenir 74 %, une carte de Kohonen 69.9 %. Globalement les performances restent inférieures à celles obtenues avec la durée des voyelles. Est-il possible de combiner ces deux sources d'information pour distinguer ces catégories grammaticales ?

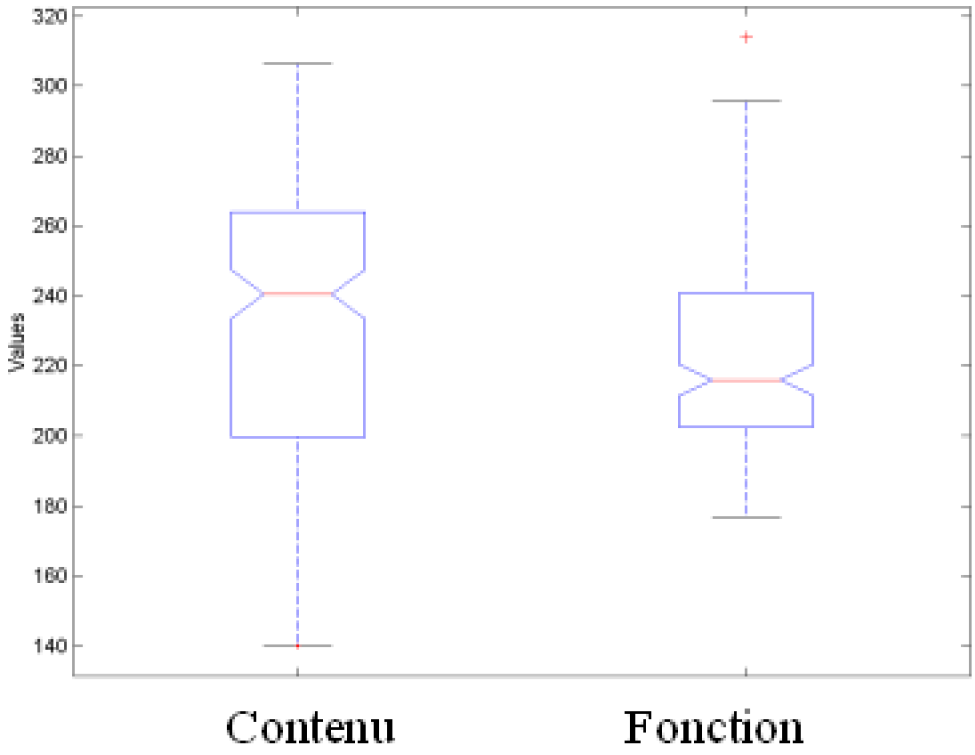


Figure 5.3 Répartition des valeurs moyennes de F0 sur la dernière voyelle d'un groupe de mots.

IV.1.1.3.Combinaison de la durée et de F0

Les groupes de mots sont caractérisés pour cette expérience par la durée et la F0 moyenne de la dernière voyelle. Ces deux paramètres sont normalisés entre 0 et 1, ainsi chaque paramètre a le même poids. La réunion de ces deux variables, et l'utilisation d'un prototype moyen autorisent un score de 72.5 %, un réseau probabiliste un score de 77.5 %, une carte auto-organisatrice (5 x 5, 2500 cycles), un score de 76.5 %.

Pour chacun des trois types d'apprentissages, les performances sont améliorées par rapport à l'utilisation du seul indice de durée de la voyelle. La combinaison de plusieurs indices peut donc être effective pour notre tâche d'analyse lexicale, comme cela a déjà fait l'objet de nombreuses constations (Shi et coll., 1998 ; Christiansen et Dale, 2001; Reali et coll., 2003).

IV.1.1.4.Discussion

Ces premiers points ont révélé deux faits :

1. Plusieurs dimensions (F0 et durée) peuvent être combinées pour distinguer les mots de fonction des mots de contenu.
2. La frontière indiquant le début d'un groupe de mot peut ne pas être prise en compte pour faire cette discrimination.

Cependant, ce travail reste limité aux groupes de mots d'un même type. Il est probable que les performances soient inférieures si les mots sont considérés, au lieu des groupes.

IV.1.2.Durée des groupes de mots

Le paragraphe précédent tenait compte de la durée des voyelles uniquement. Shi et coll. (1998) ont tenu compte de deux indices susceptibles d'induire la durée totale d'un mot. Ces deux indices sont la durée moyenne des syllabes d'un mot, et le nombre de syllabes de ce mot. Quelles peuvent être les performances de discrimination à partir de cette durée ? Les groupes de mots de fonction sont dans leur ensemble plus courts que les groupes de mots de contenu. Intuitivement, la prise en compte des mots devrait améliorer ces performances.

Le prototype moyen obtenu à partir de la durée des groupes mots permet de distinguer les deux catégories avec un score de 83.1 %. Un réseau probabiliste donne un score de 84.5 %, une carte auto-organisatrice une performances similaire de 84.5 %.

Les expériences précédentes ont montré que la durée pouvait être considérée sur les groupes de mots eux-même, sans considérer directement les voyelles. Est-il possible de caractériser la fréquence fondamentale sur la même durée ?

IV.1.3.Prototype de contour intonatif

Le traitement de l'information de F0 contenue dans un mot nécessite un peu plus d'opérations que le traitement de la durée. Effectivement, la fréquence fondamentale est caractérisée par une suite de valeurs, qui génèrent un contour intonatif défini sur un mot.

IV.1.3.1.Prototype de F0 pour un groupe de mots

Le graphe suivant donne une illustration du parcours moyen de la F0 pour les deux catégories grammaticales au cours du temps (Figure 5.4). Le premier point soulevé par ce graphe est la prise en compte de la durée. Les mots de contenu sont plus longs que les mots de fonction, et leur voisement également. En voulant caractériser l'évolution de F0, il semble impossible de ne pas tenir compte des frontières de mots.

Ce graphe (Figure 5.4) suggère que l'évolution de la F0 est différente suivant les types des mots. Le trajet pris par la F0 est beaucoup plus court pour les mots de fonction, et est donc moins sinueux que pour les mots de contenu.

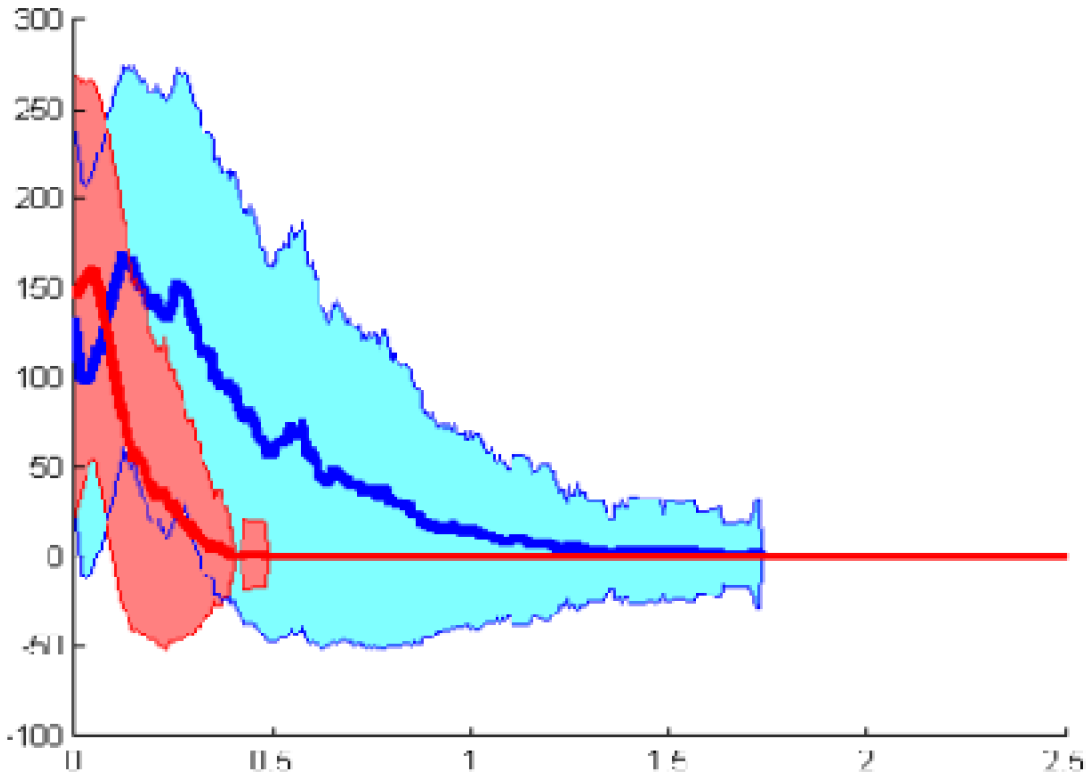


Figure 5.4 Décours temporels moyen pour la F0 (plus ou moins l'écart type pour les courbes fines) en fonction des groupes de mots de fonction (en rouges) et de contenu (en bleu).

Notre objectif est donc de pouvoir décrire l'évolution de la F0 au cours du temps à l'aide d'un nombre réduit de paramètres. Les premiers paramètres étudiés pour l'établissement d'un prototype moyen de F0 pour chacune des deux catégories sont énumérés dan le tableau 5.2.

Tableau 5.2 Performances d'identification par réseau probabiliste pour divers paramètres.

Durée des groupes de mots	84,5 %
Première valeur de F0	66 %
Dernière valeur de F0	80,6 %
Variation de F0 (première moins dernière valeur)	72,6 %
Maximum de F0	62,1 %
Position du maximum de F0	85,4 %

L'étude des indices semble une fois de plus donner raison à l'hypothèse d'intégration de dimensions différentes. Effectivement, la durée précédant le maximum de F0 permet les meilleures performances d'identification, or cet indice fait appel à la fois à la durée (le rythme) et à la fréquence fondamentale (l'intonation). L'impact de la valeur maximale peut être améliorée (70,4 %) si la valeur moyenne de F0 est également prise en considération (maximum moins la valeur moyenne sur le segment).

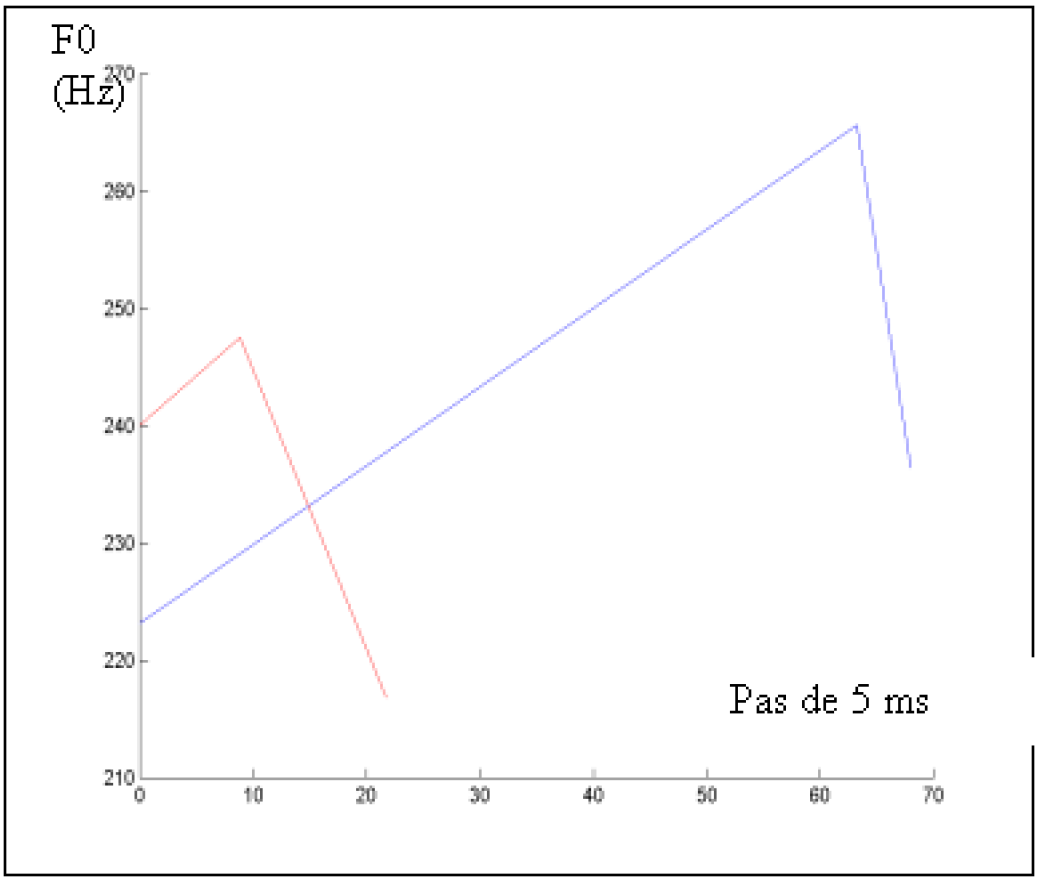


Figure 5.5 Représentation des prototypes de F0 pour les groupes fonction (pic à gauche) et contenu (pic à droite).

Empiriquement, nous avons testé plusieurs configuration des indices précédents pour obtenir les meilleures performances. La combinaison des variations, de la durée, de la valeur finale, ainsi que de la position du maximum permet d'atteindre 91 % de réponses correctes. Ces mêmes indices testés avec une carte de Kohonen permettent de retrouver des performances voisines (2500 cycles, 88.3 % pour 5 x 5 unités).

Ces expériences prouvent qu'il est possible d'identifier les mots de fonction et de contenu à partir d'un certain nombre d'indices, caractérisant l'intonation et la durée. En outre, l'intégration d'indices de plusieurs dimensions (durée et intonation) améliore les performances. Cependant, l'identification est effectuée en connaissant les frontières de début et de fin de groupes de mots. Est-il possible de ne tenir compte que de la dernière voyelle d'un groupe de mot ?

IV.1.3.2.Prototype de F0 pour la dernière voyelle

Nous cherchons maintenant un prototype de F0 uniquement pour la dernière voyelle d'un groupe. Ainsi, le début d'un mot ne sera pas pris en compte pour la paramétrisation des groupes de mots.

Nous gardons trois des quatre indices révélés précédemment (durée de la voyelle, position du maximum et variation de F0). La valeur finale de F0 ne présente pas de

différence significative entre les deux catégories syntaxiques (ANOVA, P = 0.3). Nous prendrons cette fois-ci la valeur maximale de F0 (ANOVA, P < 0.001). Nous avons dans un premier temps étudier la valeurs de probabilité de chaque indice, puis nous avons testé empiriquement les valeurs offrant les plus grandes différences. Onze valeurs statistiques de F0 ont été testées.

Un réseau probabiliste donne un taux de réponses correctes de 82,1 %. Ces performances sont inférieures à celles basées sur le segment complet, cependant il reste possible de faire la discrimination lexicale à l'aide de la dernière frontière des groupes de mots. Dans ce contexte, cette dernière voyelle peut-elle être identifiée lorsqu'elle se situe avant un mot de fonction, c'est-à-dire à la fin d'un groupe de mot contenu ?

IV.1.3.3.Discrimination de la dernière voyelle d'un groupe de mot de contenu

Les indices retenus sont les mêmes que pour l'expérience précédente, mais le but de la tâche est légèrement différent : il faut identifier la voyelle précédant un mot de fonction. Pour ce faire, nous n'emploierons donc aucune frontière, mais seul le début des groupes de mots de fonction pourront être identifiés.

Les performances sont de 80,8 % avec un réseau probabiliste. Mais les deux catégories de voyelles (précédent un mot de fonction, et autres) n'ont pas le même nombre d'éléments à identifier. En effectuant la moyenne de la reconnaissance des deux types de voyelles, le score est de 64,2 %., ce qui est supérieur à un score aléatoire (50 %). Dans ce cas, le contexte précédent les mots de fonction permet de déceler une partie de ces mots.

Nous signalions précédemment que ces études sont menées directement sur les valeurs de F0. Or ces valeurs ne sont peut-être pas accédées directement chez l'être humain. Effectivement la perception de l'intonation est différente des valeurs brutes de F0. Il serait raisonnable de tester la discrimination lexicale avec une représentation de la F0 plus proche de la perception de l'intonation.

IV.1.4.Pics de F0

La section précédente a décrit une méthode pour établir un prototype du contour intonatif. Notre objectif est maintenant de considérer un seul indice qui puisse être facilement perçu par l'oreille humaine, comme la présence d'un pic intonatif ?

Nous voulons tester l'hypothèse suivante : « La présence d'un pic de F0 indique qu'un mot appartient à la catégorie Contenu, (l'absence de cet indice indique donc un mot de fonction) ».

La localisation des pics F0 est effectuée à partir de la détection d'un changement de signe de la différence entre deux valeurs adjacentes. Cela revient à calculer les minima et maxima locaux de F0. Cependant les données brutes comportent un certain nombre de pics qui ne sont pas repérés perceptuellement.

Le chapitre 2 a introduit quelques représentations de la prosodie. L'algorithme MOMEL décrit l'intonation automatiquement à partir des valeurs brutes de la fréquence fondamentale. Nous suggérons de comparer l'algorithme MOMEL à diverses

paramétrisations de l'obtention de la F0 brute, toujours pour distinguer les groupes de mots de fonction des groupes de mots de contenu, mais pas rapport à un seul indice issu de l'intonation, la présence d'un pic. Effectivement, nos études précédentes laissent penser que la présence d'un pic, vraisemblablement représentée par la position du maximum est un indice pertinent pour cette tâche. Notre première tâche va donc consister à tester différentes façons d'obtenir des pics de F0, de façon à trouver le score maximum d'identification pour le corpus LSCP Français.

IV.1.4.1.Développement : Divers méthodes pour F0

Nous avons a priori trois méthodes distinctes pour obtenir une représentation de l'intonation. La première est de conserver les valeurs brutes de la fréquence fondamentale. En prenant ces valeurs, des pics qui ne peuvent être perçus vont apparaître. Une première solution consiste à augmenter la taille de la fenêtre d'analyse. Cependant, le nombre d'échantillon sera considérablement réduit. Une autre méthode est d'effectuer un lissage des données par l'application d'une fonction spline, après interpolation des données (opérations réalisées avec le logiciel PRAAT). La dernière opération sera l'application de l'algorithme MOMEL.

Nous testerons donc quatre méthodes :

1. Valeurs brutes d'autocorrélation (deux tailles de fenêtres) pour une fenêtre de 10 ms ;
2. Interpolation puis lissage pour une fenêtre de 10 ms ;
3. Interpolation puis lissage pour une fenêtre de 60 ms ;
4. Algorithme MOMEL.

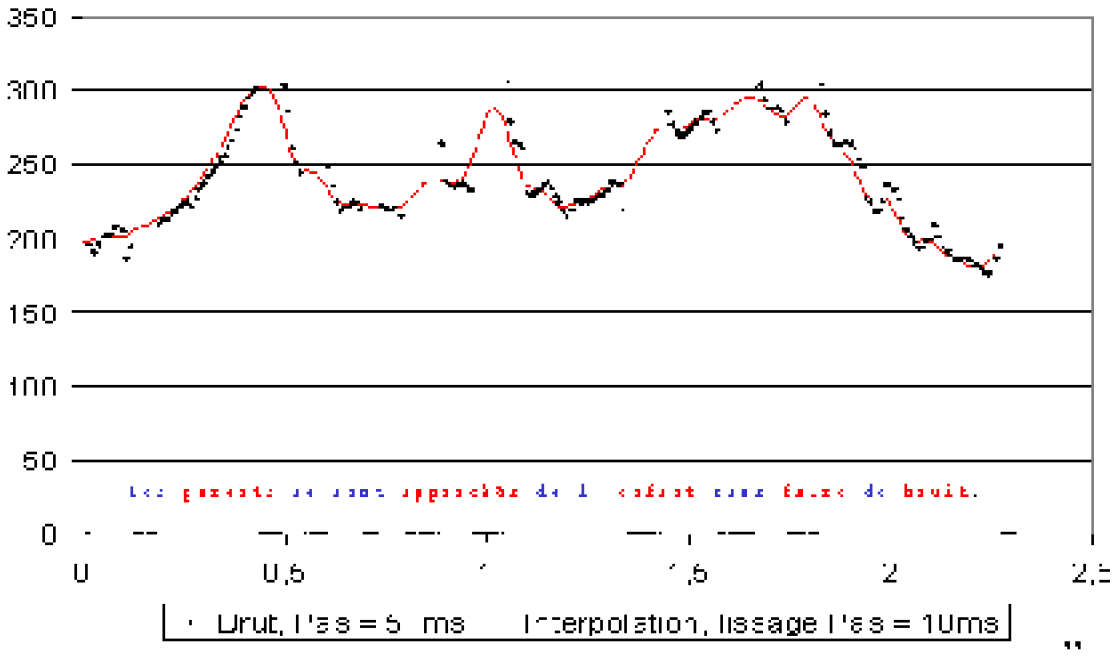


Figure 5.6 Représentation de F0 avec une fenêtre de 10 ms, lissage et interpolation.

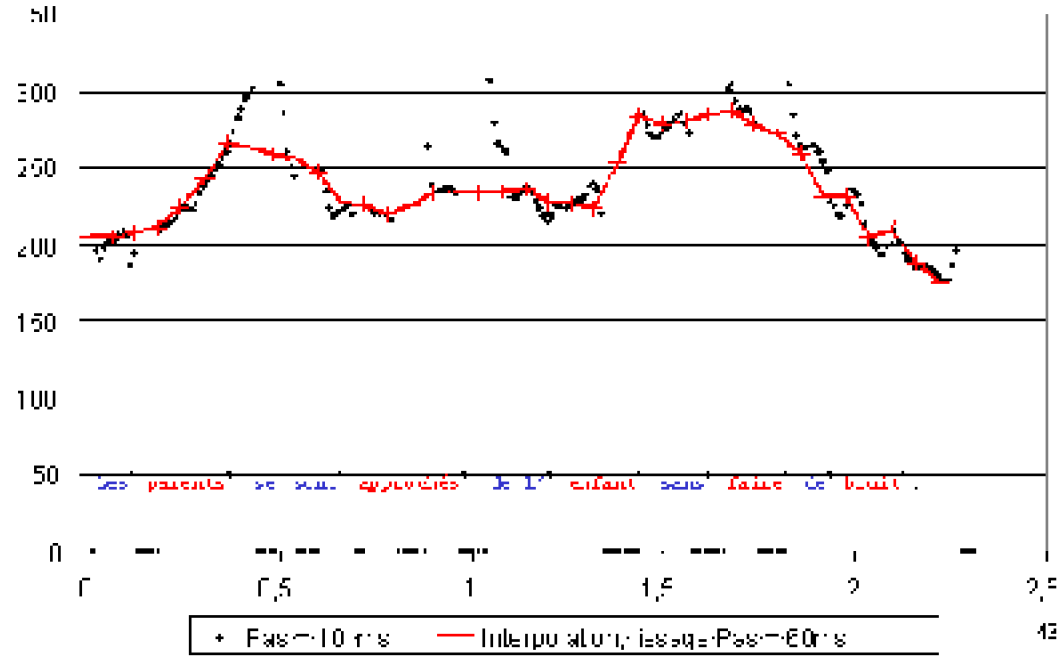


Figure 5.7 Représentation de F0 avec une fenêtre de 60 ms, lissage et interpolation. En tenant compte d'une fenêtre d'analyse aussi importante, des pics de F0 disparaissent des données.

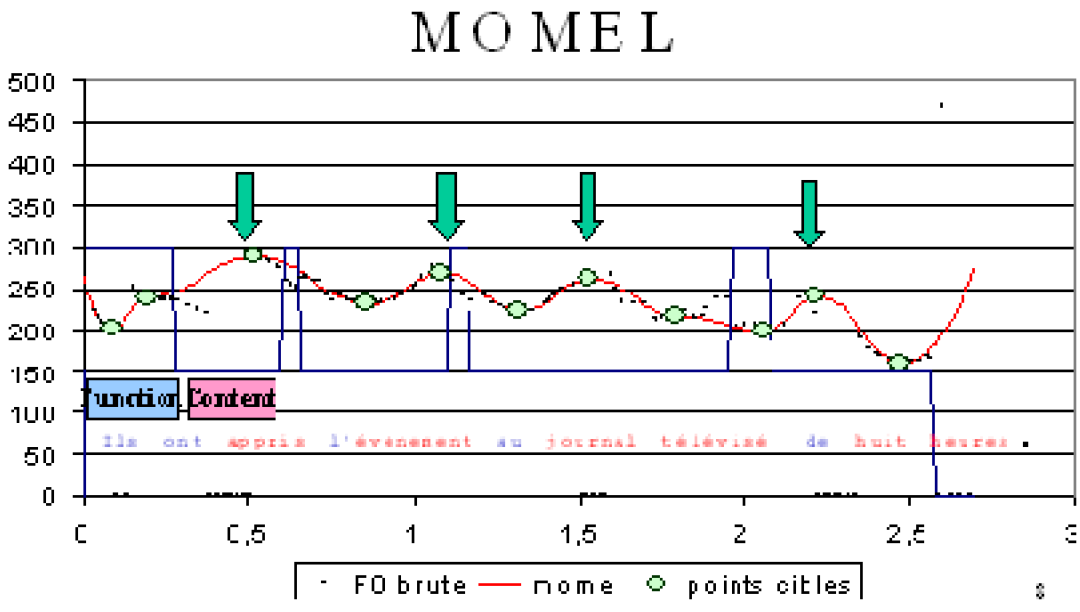


Figure 5.8 Application de l'algorithme MOMEL, les flèches indiquent la présence d'un pic de F0.

	Méthode	% correct
1	f0 brute	62,1%
2	f0 10ms	82,0%
3	f0 60ms	85,4%
4	f0 momel	76,6%

Tableau 5.3 Performance de l'identification lexicale à partir de la détection des pics pour différentes façons de représenter F0.

Le Tableau 5.3 donne le pourcentage d'identification des mots. Avec le traitement MOMEL, il apparaît que 92 % de pics F0 et 61 % de vallées F0 se sont produits dans des mots d'un contenu, avec une distribution complémentaire pour les mots de fonction.

IV.1.4.2.Discussion

Comme nous le pensions le signal formé par les valeurs brutes de F0 ne peut permettre de répondre à la tâche. Il faut au minimum lisser et interpoler les données pour pouvoir trouver les pics de F0. En employant une fenêtre plus large, l'identification lexicale est réalisée avec les meilleures performances. Cependant, il est probable que cette amélioration soit due aux indices de durées des mots. Effectivement avec des échantillons aussi espacés, certains mots de fonction ne sont pas représentés par des valeurs de F0, il devient alors impossible de former un pic sur certains mots de fonction.

En outre, la représentation donnée par MOMEL a été appréciée comme proche de la perception, puisque ces mêmes pics ont été jugés perceptibles sur le corpus MULTEXT (Campione et Veronis, 1998). Nous avons également tenté de nous passer de la frontière de début des groupes de mots. Cependant, les performances sont plutôt décevantes, il semblerait que la position du pic soit trop imprévisible, pour garder une fenêtre de taille fixe à partir de la fin d'un groupe de mots.

En moyenne les pics sont espacés de 71.5 ms pour la partie Française du corpus LSCP. A partir de cette valeur nous répartissons les pics au hasard (suivant une loi normale centrée sur l'espace moyen entre deux pics et pour écart-type, l'écart-type de ces espaces). Nous comptons ainsi vérifier quelle peut être l'influence de la durée des mots pour l'identification lexicale fondée sur la présence de pics de F0. Effectivement, les mots de fonction sont très courts, il y a donc statistiquement peu de chance qu'un pic de F0 soit sur un mot de fonction. Avec une répartition aléatoire des pics, 80 % des mots de fonction sont encore identifiés correctement, et 60 % des mots de contenus sont encore identifiés. Le score d'identification est alors de 70 %, ce qui reste inférieur aux performances données par les vrais pics de F0. Malgré tout la durée des mots de fonction (qui marque pour une grande part leur minimalité) est un facteur qui a une grande influence sur l'identification lexicale.

Nous devons noter qu'il conviendrait de traiter à part les fins de phrases, caractérisées par une descente de l'intonation ; ainsi que les continuations majeures qui peuvent avoir lieu sur des mots de fonction, et former dans ce cas un pic de F0 sur ces mots de fonction.

IV.1.4.3.Translation des pics de F0

La position des pics est liée à l'estimation de la fréquence fondamentale. Il peut arriver que les pics soit alors mal positionnés, en particulier qu'un pic tombe sur un mot de fonction, alors qu'il se trouve à la limite entre deux groupes de mots différents. Pour ce faire, nous avons translaté les valeurs de F0 (obtenues après l'algorithme MOMEL) d'un certain nombre d'échantillon (jusqu'à 100 ms de décalage). Il s'avère qu'un décalage de 100 ms permet d'accroître les performances. Cependant, il est également possible que le pic tombe intentionnellement "entre" deux types de mots différents. Dans ces cas, le pic n'indique pas la nature d'un mot, mais un changement de nature. Toutefois, l'impact reste important (+10 % de mots reconnus).

Ce décalage des pics de F0 a été déjà été signalé dans la littérature dans l'alignement des pics de F0 par rapport à des segments comme des syllabes (Silverman et Pierrehumbert, 1990 ; Prieto, van Santen et Hirschberg, 1995 ; Arvaniti, Ladd et Mennen, 1998 ; Ladd, Mennen et Schepman, 2000 ; Xu, 2001). Il semble que pour notre étude le décalage des pics sur les syllabes puisse se répercuter sur l'identification des catégories lexicales.

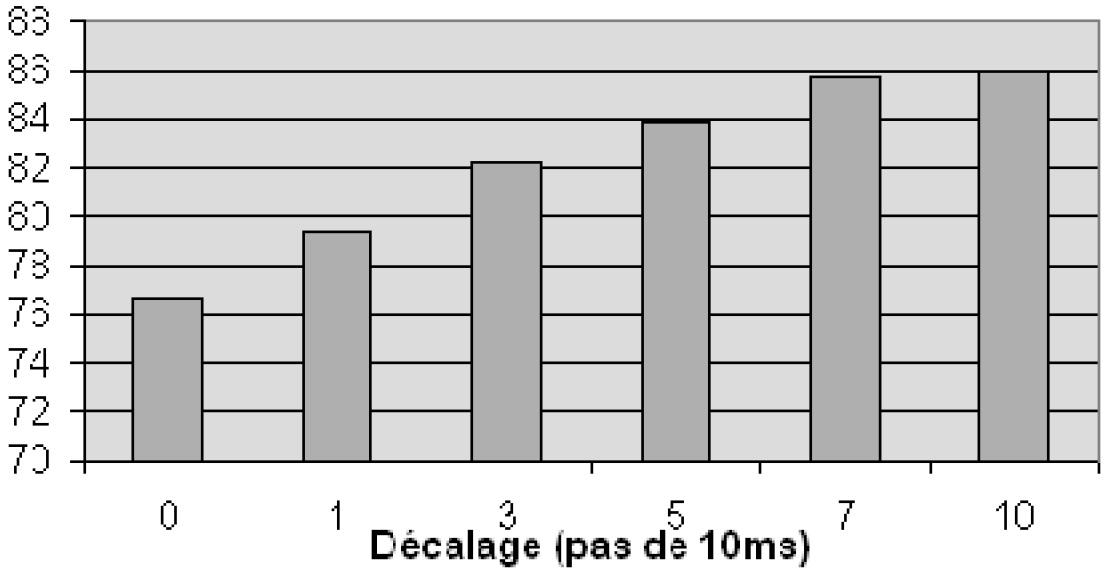


Figure 5.9 Effet de la translation des pics sur l'identification lexicale.

IV.1.4.4.Extension à d'autres corpora et aux mots

L'identification des mots de fonction et de contenu du corpus LSCP à partir des pics de F0 donne les performances, résumées dans le tableau 5.6. Un tirage aléatoire entre les deux catégories conduirait à un score d'identification de 50.6 %.

L'intérêt de la représentation MOMEL est accru par la possibilité d'utiliser le corpus MULTEXT, qui a permis de juger la validité de cette représentation. En outre, ce corpus est segmenté en mots, ce qui nous autorisera à comparer les performances entre les

groupes de mots et les mots, pour une autre langue que le Français.

Corpus	MULTEXT	MULTEXT
Langues	Anglais	Français
Pics	76%	82%
Vallées	69%	60%

Tableau 5.4 Pourcentage de pics et de vallées dans des mots de contenu.

Ce tableau 5.4 reflète déjà une différence entre l'Anglais et le Français. Dans le cas du Français, un plus grand nombre de pics sont présents, alors que moins de vallées apparaissent dans les mots de contenu, laissant présager que celles-ci seront sur les mots de fonction.

	groupes	mots
Anglais	70,7%	64,5%
Français	75,1%	73,1%

Tableau 5.5 Performances de la détection des pics pour le corpus MULTEXT pour des groupes de mots et les mots eux-même avec l'algorithme MOMEL.

	groupes	mots
Anglais	75,5%	62,1%
Français	76,6%	65,9%

Tableau 5.6 Performances de la détection des pics pour le corpus LSCP pour de groupes de mots et les mots eux-même avec l'algorithme MOMEL.

Au premier abord, les performances obtenues sur le Français avoisinent celles du corpus précédent, et le passage aux mots dégrade peu les performances. Les résultats sont différents pour le corpus LSCP. Les performances diminuent nettement pour les mots. Cela est peut-être du à un mauvais repérage des pics par l'algorithme MOMEL sur le corpus LSCP. Lorsque les pics sont décalés de 10 pas, nous obtenons un taux de réussite de 75 % pour les mots. Toutefois les performances pour l'Anglais suivent la même diminution pour le corpus MULTEXT.

Nous appliquons alors les méthodes d'interpolation et de lissage fournies dans PRAAT aux valeurs brutes de F0 obtenus avec une fenêtre d'analyse de 10ms. Les performances de l'identification des mots pour le corpus LSCP sont pour l'Anglais 76,6 % et pour le Français 86,3 %. Ceci indique que les pics de F0 donnés par ces méthodes sont utiles pour l'identification des mots de fonction et de contenu. Les performance sont caractéristiques des différences observées entre l'Anglais et le Français. Cependant, l'application de ces méthodes ne garantit pas que les pics de F0 détectés seraient perceptibles par l'oreille humaine.

Pour le corpus MULTEXT Anglais, une perte de 6 % (2 % en Français) est observée entre les mots et les groupes. Cette différence illustrent le fait qu'un même indice peut avoir un impact différent pour l'identification lexicale, comme cela a déjà été souligné par Shi et coll. (1998) et Morgan et coll. (1996). En outre, la détection des pics est

indépendante du locuteur, puisque celle-ci ne dépend pas de la fréquence moyenne, mais est déterminée par le contexte.

Nous n'avons tenu compte pour l'instant que de la fréquence fondamentale et de la durée. Des dimensions acoustiques supplémentaires, telles qu'intensité et formants peuvent-elles être employées pour distinguer mots de fonction et mots de contenu ?

IV.1.5.Prototypes prosodiques

IV.1.5.1.Création des prototypes prosodiques

Dans cette section, nous allons étendre les travaux concernant les prototypes de F0, au corpus MULTEXT, et intégrer un plus grand nombre de dimensions prosodiques : F0, amplitude, F1, F2, F3, et les variations de F1, F2 et F3, soit 8 dimensions présentées en fonction du temps. En outre, nous utiliserons un ensemble de 15 mesures statistiques pour chaque dimension :

1. première valeur ;
2. valeur finale ;
3. valeur maximale ;
4. position du maximum ;
5. valeur minimale (différente de 0) ;
6. position du minimum ;
7. moyenne ;
8. écart type ;
9. rapport de la durée des valeur non nulles par la durée des valeurs nulles ;
10. nombre de montée et descente (données par le changement du signe de la variation) ;
11. nombre de montée et descente divisé par la durée ;
12. moments d'ordre 2 à 5 (Le moment d'ordre 2 est en lien avec la variance, celui d'ordre 3 avec le Skewness, celui d'ordre 4 avec le Kurtosis).

Nous obtenons donc un vecteur de 120 composantes qui représente un mot ou un groupe de mots de même type lexical. Nous appliquons une analyse discriminante pour évaluer les performances de classement de ces 120 indices, et leur contribution par rapport à la durée des mots (ou des groupes de mots le cas échéant).

IV.1.5.2.Identification des mots de fonction et de contenu

Le tableau 5.7 suivant indique les performances obtenues pour une analyse discriminante (logiciel SPSS) pour accomplir la tâche d'identification lexicale à partir de la durée des segments étudiés (mots ou groupes) et des prototypes prosodiques.

	MULTEXT				LSCP
	Anglais		Français		Français
	groupes	mots	groupes	mots	groupes
durée	77,3%	85,3%	72,3%	83,7%	83,8%
120 indices	79,1%	83,4%	79,5%	85,5%	-
F0	75,3%	80,9%	74,2%	82,3%	87,7%
durée & F0					91,1%
durée & 120 indices	79,1%	84,0%	79,9%	86,2%	-
différences	1,8%	-1,3%	7,6%	2,5%	7,3%

Tableau 5.7 Performances de l'analyse discriminante des prototypes prosodiques. La différence entre les lignes durée et durée & 120 indices est indiquée dans la dernière ligne.

Ce tableau 5.7 fait apparaître une différence entre l'Anglais et le Français comparable à celle qui ressortait avec l'étude des pics de F0. L'information prosodique n'améliore les performances que pour le Français et ceux principalement pour les groupes de mots. En outre, il apparaît que la durée des segments permet une identification comparable entre les langues (> 80 %), lorsqu'il s'agit des mots. Ceci confirme le caractère minimal des mots de fonction pour l'Anglais et le Français. Toutefois, il apparaît que les mots de fonction sont un peu plus souvent regroupés en Français qu'en Anglais. Les performances diminuent moins en Anglais, lorsque les groupes sont considérés.

Le corpus MULTEXT permet de valider l'identification des mots de contenu et de fonction sur plusieurs locuteurs. Dans ce cas, chaque locuteur est testé indépendamment des autres. L'apprentissage se fait sur la moitié d'un corpus pour un seul locuteur. Les performances sont voisines pour chaque locuteur. Les différences qui apparaissent peuvent être dues à la structure syntaxique des énoncés, (4 passages pour 10 locuteurs) ou à l'interprétation prosodique des locuteurs. Des analyses ultérieures pourraient comparer les mêmes énoncés prononcés par deux locuteurs différents.

IV.1.5.3.Identification des noms et des verbes

Quelques études ont montré des différences acoustiques entre les noms et les verbes (cf. II.1.2). Cette classification a été réalisée à l'aide d'un réseau probabiliste pour le Français.

L'utilisation des prototypes prosodiques permet la distinction entre les noms et les verbes. Cependant, en comparant les performances de chaque indice pris isolément, un indice permet d'avoir une performance supérieure. Seulement cet indice s'avère différent pour chaque locuteur, autant par la dimension pris en compte que par la statistique décrivant cette dimension. Toutefois, les performances dépassent celles obtenues par la durée. En Français, les verbes ne sont pas décrits de façons minimum, contrairement aux mots de fonction.

	moyenne	maximum	minimum
1	63,1%	69,7%	58,2%
2	67,2%	76,1%	60,5%
3	59,0%		

Tableau 5.8 Performances du réseau probabiliste pour la discrimination nom/verbe (performance moyenne pour dix locuteurs, et extrêma atteints par un de ces locuteurs). Trois cas sont envisagés : 1) les 120 indices prosodiques sont retenus, 2) seul l'indice donnant la meilleure performance (pour chaque locuteur) est retenu, 3) la durée des mots est prise en compte.

IV.1.5.4.Discussion

La méthode que nous présentons ici est relativement proche de celle développée dans (Oudeyer, 2002). Un nombre important d'indices statistiques est utilisé pour caractériser des segments de paroles, une technique d'apprentissage détermine alors la meilleure combinaison pour résoudre le problème posé. Cette méthode a cependant l'inconvénient de ne pas être a priori réalisable par des méthodes connexionnistes. Cette section prouve que l'intonation peut être employée pour distinguer mots de fonction et de contenu, en particulier les pics de F0 semblent avoir une place prépondérante dans cette classification.

Cependant, un point reste en suspend : la fréquence fondamentale ne correspond pas totalement à l'intonation telle qu'elle est perçue par l'être humain Est-il possible d'extraire ces indices avec un système vérifiant des recherches en neuroscience?

IV.2.Le réseau TRN

Les expériences menées avec le TRN s'orientent suivant trois axes :

1. Nous étudierons le comportement du réseau sur le corpus LSCP pour le Français et pour des groupes de mots ;
2. La représentation de F0 sera modifiée, de façon à évaluer son influence sur le réseau TRN lors de l'identification des mots ;
3. Enfin, nous testerons le réseau TRN pour identifier des mots provenant de plusieurs locuteurs (corpus MULTEXT), ainsi que pour l'Anglais et le Français (corpora MULTEXT et LSCP).

Contrairement aux méthodes précédentes, un mot sera catégorisé uniquement en tenant compte de la dernière frontière, et non plus des deux frontières encadrant le mot. Dans ce contexte, la durée des mots ne peut être détectée par le réseau TRN.

IV.2.1.Catégorisation lexicale du corpus LSCP

Cette première section est dédiée aux tests de diverses méthodes pour réaliser l'identification des mots de fonction, avec le réseau TRN. Sauf indication contraire, nous

tiendrons compte de la segmentation en groupe de mots, et des valeurs de F0 données par BLISS.

Le réseau TRN a d'abord été appliqué sans modification pour un premier test. Trois techniques ont alors été appliquées au réseau TRN. Les valeurs de F0 ont été recalculées à l'aide du logiciel PRAAT, les mots ont également été considérés à la place des groupes de mots. Enfin, deux expériences complémentaires ont été menées pour entamer une discussion sur le comportement du réseau.

IV.2.2.Premier test

Les vecteurs issus du traitement par le TRN de la première moitié des groupes de mots de fonction et des mots de contenu constituent l'apprentissage. Après présentation d'un groupe de mots, le vecteur de l'activité dans State / $State_D$ a été prélevé.

Deux types d'informations ont été transmises au réseau TRN pour représenter les groupes de mots :

1. CV illustre le rythme donné par la succession des consonnes et des voyelles ;
2. F0 la fréquence fondamentale du locuteurLes valeurs de F0 ont été obtenues à partir du logiciel BLISS (Ramus, 1999). Quelques valeurs sont probablement erronées (> 400 Hz) et dues à certains consonnes., représentée par 15 neurones.

Le réseau récurrent temporel opère une identification entre les mots de fonction et de contenu d'environ 80 % à partir de la fréquence fondamentale. Les représentations combinant la F0 avec les catégories des phonèmes ne semblent pas bien adaptées au problème. Ainsi, les performances atteintes pour le rythme seul sont voisines du hasard. Le meilleur taux d'identification étant réalisé avec la F0 seule, nous ne tiendrons plus compte du rythme induit par les phonèmes (CV).

Les valeurs données pour le rythme (CV) sont très proches de celles obtenues en donnant une entrée constante au réseau. Le seul fait de segmenter après chaque groupe permet une identification d'environ 60 %. Effectivement, le réseau est initialisé à chaque début de phrase, si bien que les mots de fonction débutant la phrase sont facilement identifiés parce qu'ils sont représentés par un vecteur nul (soit 10 % de réponses correctes supplémentaires). Si le réseau ne l'est pas, les performances sont au niveau du hasard (soit 50 %).

IV.2.2.1.Amélioration des performances

Deux hypothèses ont été testées pour augmenter le taux d'identification. La première modifie l'architecture du réseau. Puis, plusieurs méthodes d'apprentissage des motifs produits par le TRN ont été testées.

IV.2.2.1.1.Influence de la couche d'entrée du réseau

La première manière d'augmenter les performances du réseau consiste à tenir compte de différents nombres de neurones pour la fenêtre représentant la fréquence fondamentale,

et différentes valeurs de l'écart type (paramètre sigma), qui agit sur le nombre de neurones actifs pour une fréquence donnée. Ceci a été rendu possible par la construction d'un programme écrit en C++ conçu de manière à garder un choix dynamique de la taille et de l'architecture du réseau.

Le graphique présenté à la Figure 5.10 illustre les valeurs moyenne pour des population des 50 réseaux. Les performances moyennes sont améliorées pour 60 neurones avec un nombre de neurones actifs relativement élevés. Cependant, la performance maximale (82 %) reste la même. Les résultats obtenus sont donc plus robustes pour la population de réseaux. Dans ce cas, l'augmentation du nombre de neurones d'entrées permet d'accorder moins d'importance à l'ajustement des poids du réseau. Est-il possible d'améliorer ces résultats en utilisant une technique d'apprentissage plus performante que celle utilisant les prototypes ?

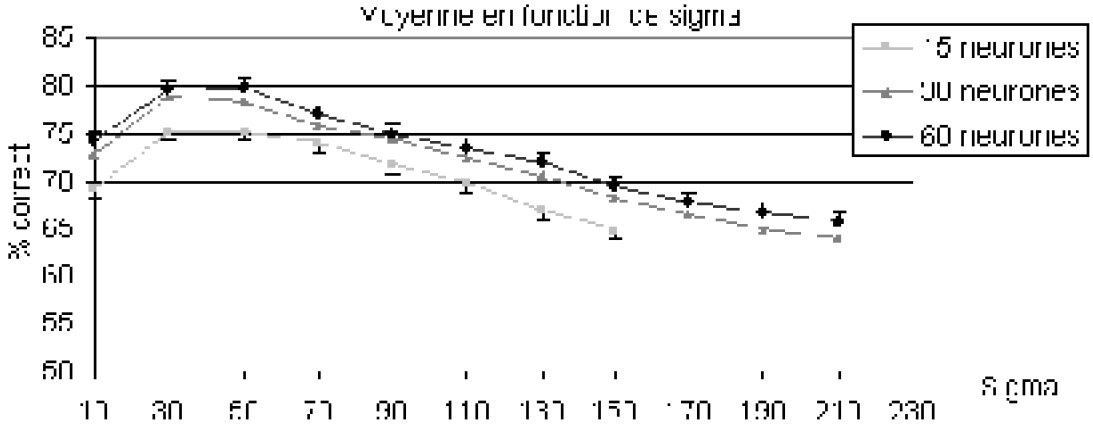


Figure 5.10 Performance d'une population de 50 réseaux en fonction du nombre de neurones d'entrées pour F0 et le paramètre sigma. Les barres verticales indiquent l'écart type des performances de la population de réseaux.

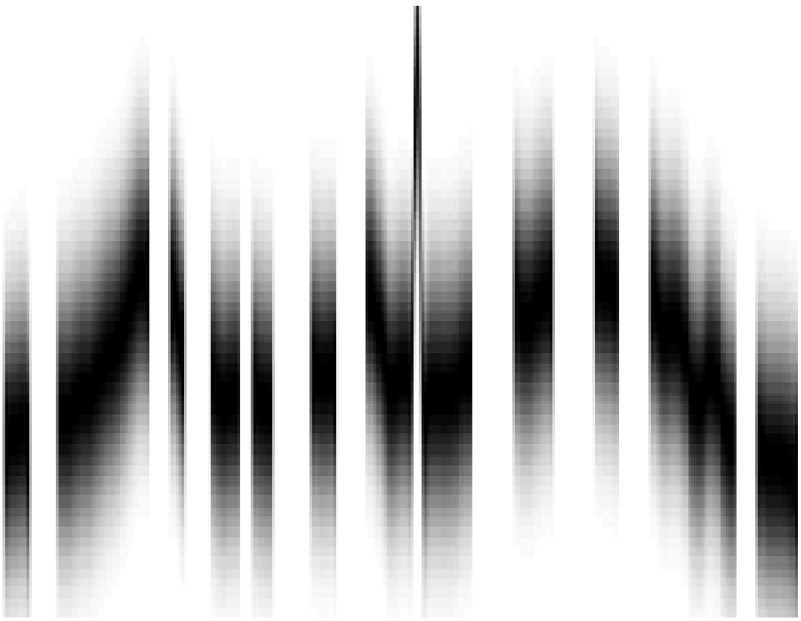


Figure 5.11 Codage retenu pour la F0 (exemple avec 60 neurones et sigma égal à 50)

IV.2.2..1.2.Différentes méthodes d'évaluation du réseau

Le tableau 5.9 donne les performances en validation pour le réseau le plus performant parmi une population de 50, avec différentes méthodes d'évaluation du réseau : 1. prototype (moyenne des vecteurs, puis distance euclidienne), 2. Analyse discriminante (Logiciel SPSS), réseau probabiliste (MATLAB : 3. soit deux corpora (apprentissage et validation) avec une sélection en validation ; 4. soit trois corpora avec une sélection effectuée sur un corpus de développement avant de pratiquer le test de validation vraiment en aveugle), 5. Carte de Kohonen (méthode d'apprentissage non supervisée, MATLAB), 6. Perceptron simple, 7. Perceptron avec deux couches cachées (10 et 5 neurones).

		Validation	
		Moyenne	Maximum
1	prototype	79,7%	83,6%
2	réseau probabiliste	83,2%	87,2%
3	réseau probabiliste (3 corpus)	78,0%	82,3%
4	analyse discriminante	80,9%	86,3%
5	Som	79,3%	84,1%
6	perceptron	76,7%	84,5%
7	perceptron (2 couches cachées)	75,8%	84,1%

Tableau 5.9 Performance de différentes méthodes d'évaluation du réseau TRN.

Ces résultats montrent que l'encodage effectué par le réseau est robuste pour diverses méthodes. Globalement, les performances maximales sont supérieures à 82 %, et les moyennes dépassent 75 %. L'analyse discriminante (logiciel de statistique SPSS) donne les meilleurs résultats, mais ils restent assez proches de l'apprentissage par prototype. Les performances des réseaux avec rétropropagation du gradient devraient vraisemblablement être améliorées en testant un plus grand nombre d'architectures.

IV.2.2.2.Nouvelles valeurs de la fréquence fondamentale et segmentation en mots

Les valeurs de F0 ont été obtenues avec le logiciel PRAAT par une technique d'autocorrélation. Nous avons aussi tenu compte d'une segmentation en mots, et non plus d'une segmentation en groupe de mots d'une même catégorie lexicale (fonction ou contenu).

Utiliser une segmentation en mots au lieu d'une segmentation en groupes ne perturbe pas vraiment les performances. En revanche, l'obtention de F0 influe plus sérieusement sur les résultats. Il est probable que les valeurs données par PRAAT contiennent moins d'erreurs, i.e. les valeurs extérieures au contour intonatif sont rejetées. Effectivement, la nature des phonèmes peut « biaiser » certaines valeurs de F0. Nous en concluons qu'une partie de l'information phonotactique transparaît au travers des valeurs de F0. Ce taux varierait suivant la méthode de calcul de F0.

		Validation	
Segmentation	F0	Moyenne	Maximum
Groupes	LSCP	79,7%	83,6%
Groupes	PRAAT	73,2%	80,5%
Mots	PRAAT	74,3%	79,5%

Tableau 5.10 Performance du TRN (méthode du prototype) suivant une segmentation en groupe de mots ou en mots et pour la F0 obtenue avec le logiciel PRAAT ou BLISS.

Tous les résultats tiennent compte uniquement des performances obtenues pour la validation sur la seconde moitié du corpus LSCP. Compte tenu de la faible taille des données, nous avons calculé les performances d'apprentissage et de validation, pour

chacune des moitiés du corpus LSCP. La moyenne de ces valeurs est de 72,8 % en apprentissage et de 71,9 % en validation. De plus, nous constatons une différence de performances entre les deux moitiés du corpus LSCP.

IV.2.2.3.Analyse du traitement effectué par le réseau

Ces expériences doivent répondre à deux interrogations :

1. Les motifs représentant les mots dans le réseau TRN peuvent-ils être appris à partir d'un petit nombre d'exemples ?
2. Quelle partie de l'information du contour de F0 est nécessaire pour identifier les mots ?

IV.2.2.3.1.Influence du nombre de mots

L'influence du nombre de mots compris dans la base d'apprentissage a été étudiée. Pour que le réseau puisse être utile pour l'amorçage des catégories syntaxiques, il est important que les prototypes puissent être construits à partir d'un nombre peu important de mots. La figure 5.12 montre que dix mots suffisent pour obtenir une représentation stable du point de vue des performances. L'opération d'amorçage prosodique peut donc se faire dès qu'une segmentation en mots est disponible, et ce, avec très peu d'exemples.

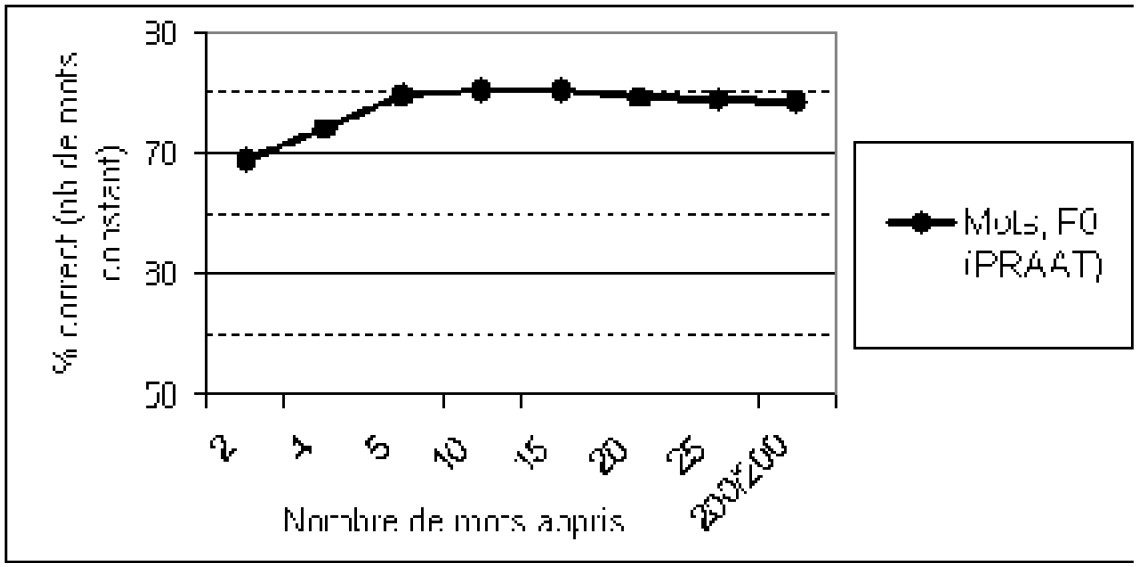


Figure 5.12 Performance en fonction du nombre de mots appris. Les performances sont indiquées pour la segmentation en mots seuls avec les valeurs de F0 obtenues avec PRAAT. Seules les performances du corpus de validation (seconde moitié du corpus LSCP) sont indiquées.

IV.2.2.3.2.Fenêtre d'analyse

Il est souvent difficile de déterminer la manière dont un modèle connexionniste apprend des données. Nous voulons comprendre d'où provient l'information utilisée par le réseau dans le signal de la F0. Pour ce faire, l'évolution de la F0 a été remplacée par une valeur

constante (la valeur moyenne de 230 Hz) pour un nombre variable de pas de la simulation. Ainsi, l'information de la F0 est dévoilée progressivement à l'intérieure d'une fenêtre à partir de la fin des mots. Les trois dernières valeurs de F0 de chaque mot sont suffisantes pour approcher le score final (Figure 5.13). Mais la totalité du segment influence quand même le score final, dans la mesure où les performances fluctuent lorsque toute l'information de F0 est dévoilée.

En utilisant un réseau probabiliste avec, en entrée, les 6 dernières valeurs de F0 de chaque segment, le taux d'identification atteint 81,2 %. En utilisant une représentation lissée et interpolée des données de F0, les performances sont moindres 76,2 %.

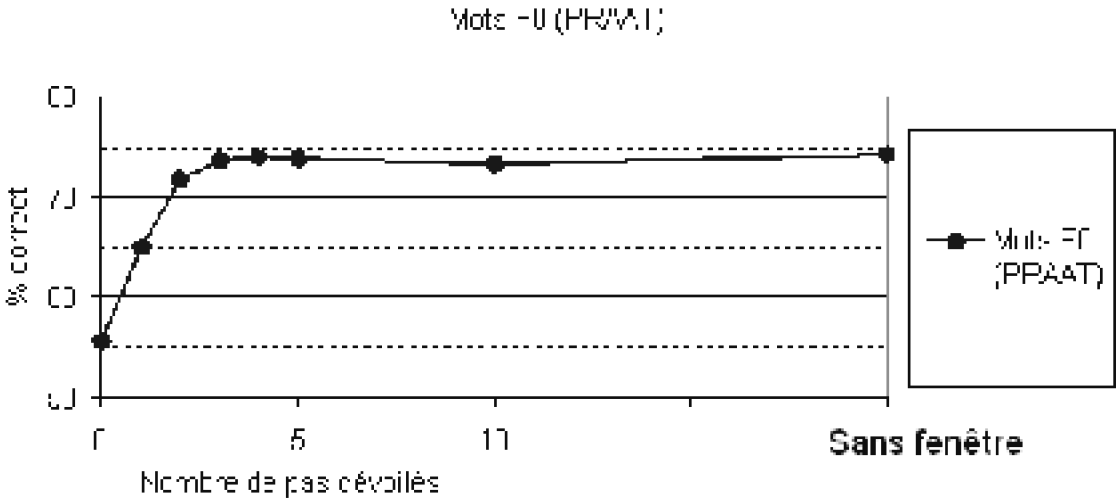


Figure 5.13 Performance en fonction du nombre de pas de F0 disponibles (moyenne pour 50 réseaux). Les performances sont indiquées pour la segmentation en mots avec les valeurs de la F0 obtenues par PRAAT.

Les indices utilisés par le TRN pourraient être liés à des variations très rapides (inférieures à 100 ms) pour caractériser les mots de contenu par rapport aux mots de fonction. Les scores diminuent lorsque des représentations qui ne possèdent pas de contrastes micro-prosodiques sont employées (F0 lissée et interpolée). En outre, cela confirme que le réseau TRN exploite plusieurs indices dans le signal de F0. Ces résultats supposent que le réseau TRN emploie certains indices microprosodiques de la F0, pour executer la tâche d'identification lexicale.

IV.2.3.Représentation spectrographique de la F0

Nous voulons maintenant estimer si le réseau TRN peut traiter une information traduite directement de manière analogique, sans passer par l'artifice de la courbe de Gauss, pour transmettre des valeurs analogiques, à partir des valeurs numériques d'autocorrélation.

La première partie consacrée à l'Identification Automatique des Langues a nécessité une représentation acoustique du signal de parole. Une répercussion de ce travail est l'utilisation d'un spectrogramme (basé sur une échelle de perception ou non) pour représenter la partie prosodique dédiée à la fréquence fondamentale. Cette nouvelle représentation n'a pas été choisie pour améliorer les performances, mais pour procurer

une représentation du signal, plus proche de l'analyse de la cochlée. Ainsi, la première couche d'entrée du réseau est constituée par une représentation spatio-temporelle du signal.

Nous avons envisagé d'étudier trois représentations des fréquences entre 0 et 400Hz :

1. un spectrogramme à bande étroite (Figure 5.14 droite) ;
2. un cochléogramme ;
3. un spectrogramme fondé sur une échelle Mel (PRAAT : Melfilter, Figure 5.14 gauche).

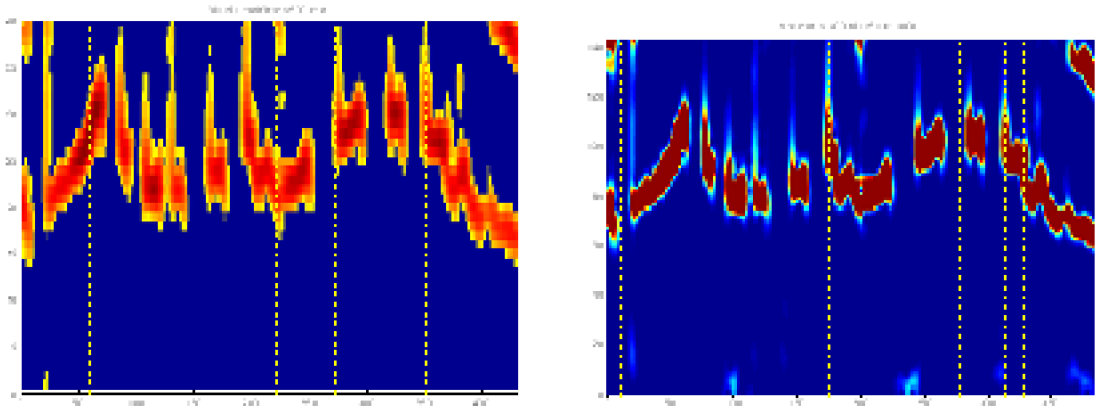


Figure 5.14 Représentation de la prosodie, à gauche par des filtres Mel (suppression des valeurs inférieures à 35), à droite par un spectrogramme à bande étroite.

Dans un premier temps, chacune de ces représentations a été évaluée sur la tâche d'identification lexicale avec le corpus LSCP (Figure 5.15). Le cochléogramme donne un résultat plutôt décevant [100] . L'utilisation de la totalité du spectrogramme à bande étroite donne des résultats honorables, ce qui confirme qu'il est possible d'utiliser la totalité du spectre pour avoir une estimation de la fréquence fondamentale. La restriction de ce spectre aux bandes de fréquences inférieures à 400 Hz donne les meilleurs résultats, proches de ceux observés avec la F0 obtenue par autocorrélation. L'utilisation des filtres basés sur une échelle Mel nécessite 40 neurones pour donner les meilleurs résultats, qui reste tout de même inférieur à la présentation de la prosodie par le spectrogramme.

[100] Ce résultat est surprenant puisque cette représentation est censée accorder une place privilègiée aux basses fréquences, contrairement à un spectrogramme.

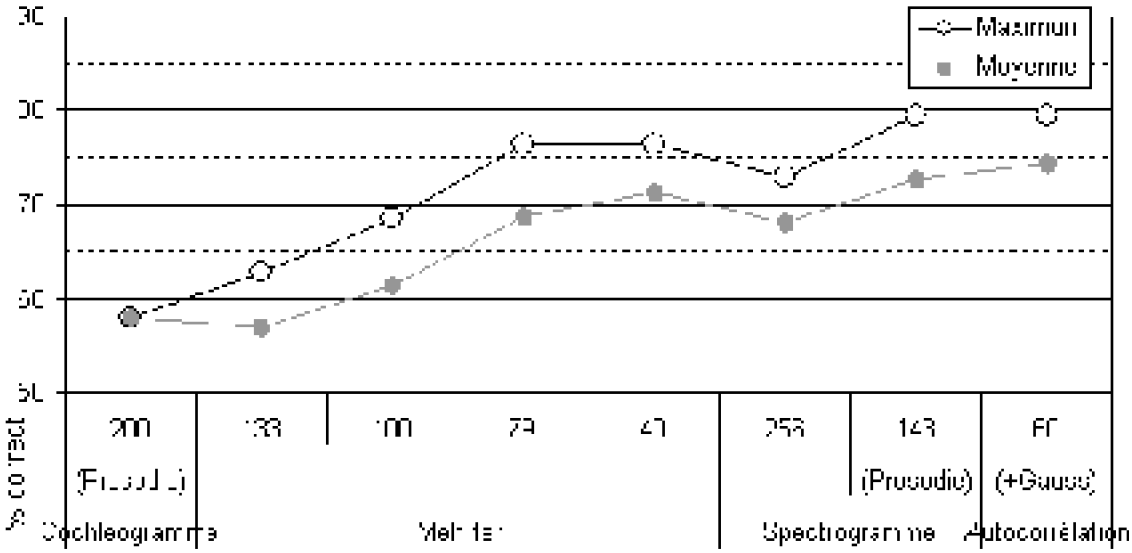


Figure 5.15 Performance pour des représentations spatiales des fréquences obtenues directement à partir du signal (corpus LSCP et logiciel PRAAT).

La différence entre la représentation Melfilter et le spectrogramme provient sans doute d'un niveau de bruit plus élevé. Effectivement la représentation par spectrogramme est beaucoup moins bruitée. Nous proposons d'ajouter un filtre qui mettra à zéro les unités dont l'intensité est faible (i.e. inférieure à seuil donné Figure 5.16).

Lors de l'étude de cette représentation par spectrogramme, deux problèmes s'opposent : 1) le réseau TRN a besoin d'un nombre important de neurones activés pour coder la F0 ; 2) plus le nombre de neurones activés est grand, plus le signal sera bruité.

L'ajout d'un filtre pour éliminer le bruit permet d'améliorer les performances (Figure 5.17). Cependant, il serait plus satisfaisant d'avoir un filtre adaptatif à l'environnement ambiant.

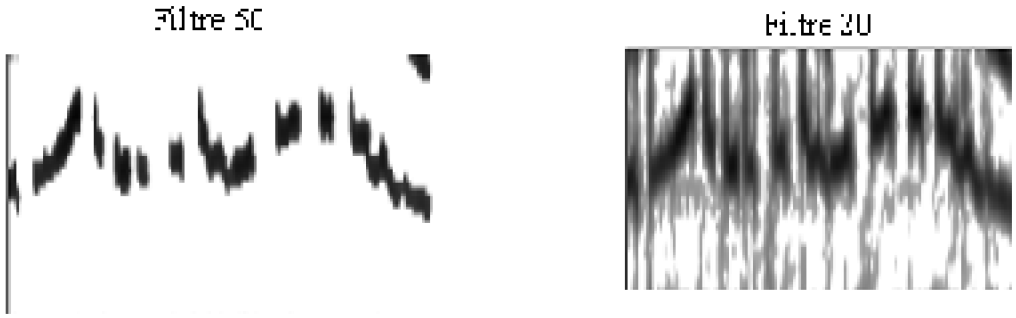


Figure 5.15 Effet d'un filtre (suppression des valeurs inférieures à 50 ou 20) sur la représentation par filtre fondée sur une échelle Mel.

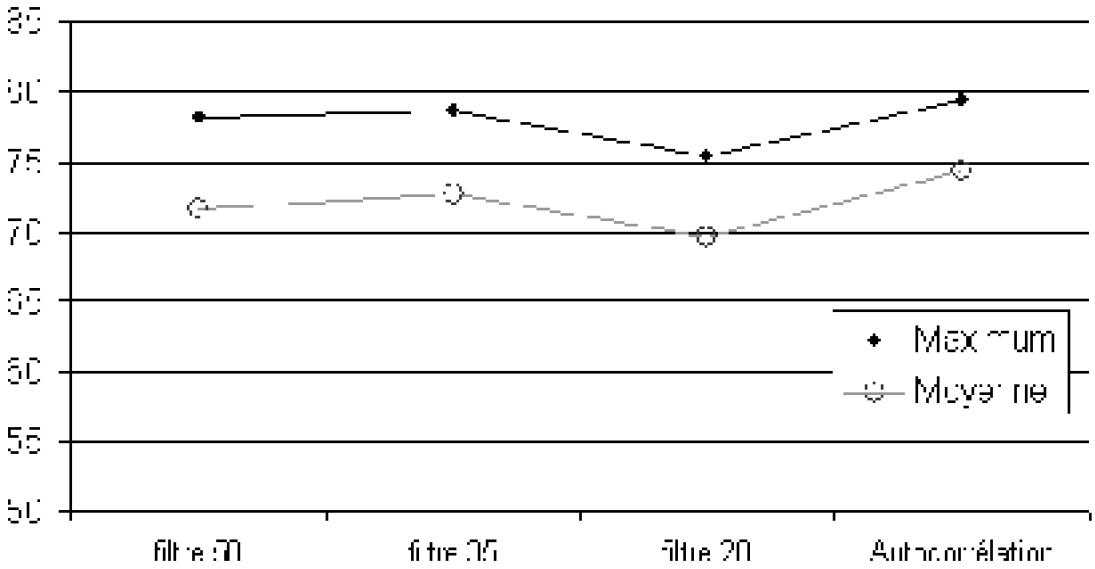


Figure 5.17 Performance pour différents seuils (50, 35, 20) de filtre (représentation melfilter).

IV.2.4.Application du TRN à d'autres langues

Nous allons maintenant examiner nos méthodes avec un corpus plus important (MULTEXT). Il est constitué de courts passages au lieu de phrases, et contient plusieurs locuteurs. En outre, le discours est plus proche d'un discours non contraint que le corpus LSCP. Il s'agit maintenant de passages et non plus de phrases isolées. En outre, nous ne présentons pas les résultats obtenus pour F0, mais ceux fournis pour une combinaison de la F0 et de l'intensité [101] . Le passage d'une représentation à l'autre ne provoque pas une différence significative.

Nous retrouvons des performances inférieures avec le corpus MULTEXT pour le Français, par rapport au corpus LSCP. Le taux d'identification de l'Anglais reste inférieur à celui du Français (Tableau 5.11 et 5.12). Comme pour les travaux précédents, nous retrouvons que le contour de la fréquence fondamentale a moins d'impact sur l'identification des mots de fonction et de contenu en Anglais qu'en Français. En outre, les performances sont indiquées pour la population de 50 réseaux. L'écart type des performances de la population est très faible, et une sélection effectuée pendant la phase d'apprentissage ne permet pas d'améliorer les performances. Il s'en suit que tous les réseaux TRN exhibent des performances voisines, et sont en mesure de distinguer les mots de contenu des mots de fonction.

	Anglais		Français	
LSCP	60,6%	(58 % ; 2,1%)	71,9%	(65 % ; 2,7%)
MULTEXT	59,9%	(53% ; 1,2%)	66,1%	(63% ; 2,1 %)

Tableau 5.11 Moyenne des performances des 50 réseaux pour l'identification des mots de fonction de contenu à partir de la fréquence fondamentale, lors de la phase de validation.

[101] L'intensité indique l'activation du neurone correspondant à la F0 calculée.

Pour le corpus LSCP, la moyenne des performances obtenues sur chaque moitié du corpus LSCP est indiquée. Pour le corpus MULTEXT, la moyenne des 10 locuteurs est donnée. Entre parenthèses, figurent le pourcentage de la catégorie majoritaire et les écart types de la population.

En outre, nous avons étudié les résultats du cochléogramme. Les indices prosodiques peuvent être utilisés pour marquer une distinction entre les mots de contenu et les mots de fonction. L'adjonction d'une représentation spectrale permet d'améliorer les performances. Le taux d'identification est inférieur pour l'Anglais (cf. Tableau 5.12).

	Anglais	Français
F0+AMP	62,8%	70,3%
Cochléogramme	62,6%	66,7%
F0+AMP+Cochléogramme	65,0%	73,6%
Pics de F0 (sans TRN)	64,5%	73,1%

Tableau 5.12 Performance du réseau le plus performant parmi 50 en validation pour l'identification lexicale des mots.

La section précédente (IV.2.1) a montré une différence de comportement suivant la segmentation (soit les groupes de mots de même catégorie lexicale, soit les mots). Effectivement, les performances sont améliorées pour les mots avec la durée, mais diminuent lorsqu'il s'agit de prototype prosodique. Le TRN utilisant la prosodie, les taux d'identification devraient donc être inférieurs dans le cas des mots.

Pour obtenir l'avant-dernière ligne du tableau 5.12, toutes les combinaisons des réseaux TRN ont été étudiées (50 x 50 essais). A titre de comparaison, nous avons effectué la même opération avec seulement les réseaux qui ont encodé la F0. Les performances sont alors de 71,5 % pour le Français, ce qui reste inférieur au mélange obtenu avec la cochlée (73,6 %). Le cochléogramme a donc une influence sur les performances d'identification. La figure 5.18 représente l'activation moyenne au cours du temps des neurones d'entrées en fonction des bandes de fréquences, pour les catégories fonction et contenu. L'intensité apparaît comme étant plus faible pour les mots de fonction pour toutes les bandes de fréquences, notamment pour celles les plus basses, correspondant à la prosodie. L'influence de cet indice devrait être étudiée de manière isolée pour l'identification lexicale.

Le TRN exécute la distinction Contenu/Fonction avec un taux proche de la détection explicite des pics F0. Ces deux taux sont supérieurs à une estimation aléatoire. Les corpora Français LSCP et MULTEXT donnent des performances différentes, mais leur contenu syntaxique diffère également. Le corpus LSCP contient seulement des phrases entre 15 et 21 syllabes, alors que le corpus MULTEXT est constitué de courts passages, ayant une structure syntaxique plus proche du discours spontané. Néanmoins ces distinctions ne rendent pas impossible la catégorisation lexicale.

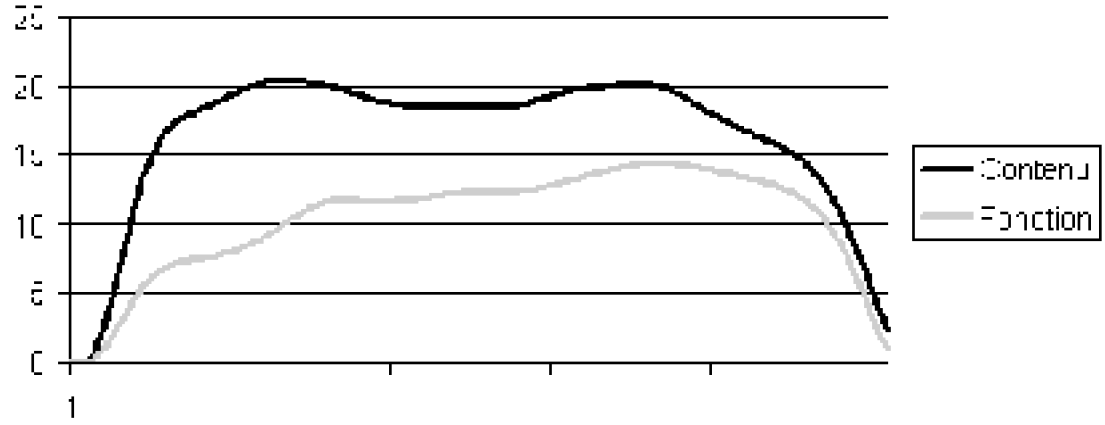


Figure 5.18 Moyenne des activations spectrales pour les mots de fonction et de contenu, en abscisse les fréquences et en ordonnées la valeur moyenne d'activation (obtenue à partir du cochléogramme sur le corpus MULTEXT pour le Français)

L'intérêt de ce travail est également d'observer le comportement du TRN pour d'autres langues. Le tableau 5.13 indique que le contour intonatif peut également être employé pour identifier les catégories lexicales pour l'Anglais. En conclusion, le TRN peut exectuer une catégorisation lexicale des mots isolés comme les nouveau-nés ont pu l'accomplir dans la tâche perceptuelle décrite dans Shi et coll. (1999).

		Contenu	Fonction
LSCP	Anglais	53,3%	69,6%
	Français	70,3%	73,7%
MULTEXT	Anglais	52,7%	68,4%
	Français	64,9%	66,9%

Tableau 5.13 Moyenne des performances des 50 réseaux pour l'identification des mots de fonction et de contenu à partir de F0 (Moyenne des 10 locuteurs pour le corpus MULTEXT et moyenne des deux moitiés du corpus LSCP)

Le tableau 5.13 donne les performances d'identification de chaque catégorie fonction et contenu. La majorité des mots de contenu sont identifiés en Français, alors que ceux-ci sont moins correctement identifiés en Anglais. Pour les deux langues, plus de la moité des mots de contenu peuvent être identifiés. Les mots de fonction sont identifiés de façon identiques dans les deux langues.

V.Discussion

La discussion des résultats précédents s'articule autour de trois axes : les problèmes soulevés par l'identification des mots de fonction et de contenu, l'extension de ce travail à un corpus plus conséquent et à une nouvelle langue, et l'implication de nos recherches par rapport à l'hypothèse d'amorçage prosodique. Enfin nous proposerons quelques piste

de recherches susceptibles d'enrichir ce travail.

V.1.Résoudre l'identification lexicale

Nous avons abordé trois des questions liées à l'identification des mots de fonction et de contenu : l'intégration d'indices prosodiques ainsi que le traitement de ces indices provenant de sources diverses et dans une moindre mesure, la segmentation de ces mots.

V.1.1.Les indices prosodiques

Quelques articles ont étudié les indices prosodiques permettant la distinction entre les mots de contenu et les mots de fonction (Morgan et coll., 1996 ; Shi et coll., 1998 ; Durieux et Gillis, 2000 ; Monaghan et coll., 2003). Cette littérature stipule que de nombreux indices acoustiques et phonologiques peuvent être utilisés par diverses approches automatiques pour distinguer les mots de contenu et les mots de fonction. Nous avons retrouvé que la durée des voyelles est un indice important (*cf.* IV.1.1). En outre, nous avons prouvé que la fréquence fondamentale contenait des informations susceptibles d'améliorer cette distinction, mais plus particulièrement dans le cas du Français (*cf.* IV.1.2, IV.2.1 et IV.2.2).

Ensuite, nous avons proposé une interprétation en terme de perception de ces résultats en comparant les prototypes de la F0 à une représentation des pics de F0. En utilisant l'algorithme MOMEL qui a été démontré pour représenter les composantes macro-prosodiques de l'intonation, nous avons montré que les pics de F0 caractérisaient les mots de contenu en Français (*cf.* IV.1.4). Ceci suggère que l'intonation ne pouvait pas être représentée correctement par l'écart type des valeurs de F0 (normalisé par la durée) dans l'article de Shi et coll. (1998), étant donné que le taux d'identification est nettement supérieur au hasard dans notre cas, et ce même pour l'Anglais.

L'intonation ne semble pas avoir été examinée, pour savoir si les nouveau-nés peuvent l'utiliser pour l'identification des mots de fonction et de contenu. Etant donné la sensibilité des nouveau-nés pour les contours prosodiques, une expérience perceptuelle pourrait montrer leur aptitude à effectuer cette distinction lexicale à partir des pics de F0 .

Le fait que les mots de fonction soient rarement marqués par un pic de F0 exprime que ceux-ci sont moins saillants dans le signal de parole, ce qui rejoint l'hypothèse de « minimalité » exprimée pour les mots de fonction (Shi et coll., 1998).

Enfin, nous retrouvons également à travers d'autres indices que les mots de fonction sont minimaux. Effectivement, ces mots peuvent être distingués par leur durée, généralement plus courte que les mots de contenu. Ce dernier point nous amène à discuter les résultats proposés par Shi et coll. (1998). Ils tiennent compte dans leurs indices de la durée moyenne des syllabes ainsi que du nombre de syllabes. En conséquence ces indices combinés peuvent donner la durée totale des mots. Or, nous montrons dans notre étude que la durée des mots permet de distinguer, aux alentours de 80 %, les mots de fonction des mots de contenu. Leurs études indiquent que le meilleur indice (durée des voyelles) pris isolément a des performances moindres.

Traitement de la prosodie

Ces indices prosodiques ont pu être employés par deux méthodes : une statistique et une connexionniste.

V.1.1.1.Prototype

La première méthode repose sur l'utilisation de prototypes prosodiques formés de statistiques sur les données acoustiques du signal au cours du temps. Cette approche ressemble à celle développée dans Oudeyer (2002) pour traiter la prosodie afin d'identifier des émotions. Cependant, avec cette méthode de traitement, nous remarquons que les indices intonatifs ont peu d'impact dans leur ensemble comparés à la durée des mots dans le cas de l'Anglais. Pour identifier les mots nous restons dépendants de leur durée. En revanche, l'intonation a plus d'impact sur l'identification lexicale dans le cas du Français (+8 % pour les groupes, et 3 % pour les mots) que pour l'Anglais.

V.1.1.2.Réseau récurrent temporel TRN

Les indices présentés dans cette étude peuvent être détectés par les enfants, mais probablement pas avec la précision utilisée ici. Est-il possible qu'il existe des structures du cerveau correspondant à des instruments de mesures pouvant donner ces indices, comme un détecteur de pics de F0 ?

Notre originalité est d'utiliser un système inspiré par des expériences en neurosciences pour l'apprentissage de séquences chez les singes (Dominey et coll., 1995). Nous montrons que ce réseau (TRN) utilise des indices micro-prosodiques pour discriminer les mots de fonction des mots de contenu, mais également l'évolution de la F0 au cours du temps.

Aucun des articles précités ne prend en compte des indices isolés automatiquement. Effectivement tous les indices indiqués sont extraits manuellement par un expert de la langue étudiée. Dans notre cas, le signal de parole est traduit sous la forme des valeurs brutes de la fréquence fondamentale. Nous proposons un système susceptible de traiter cette information, sans qu'elle soit remaniée par un expert humain. De surcroît ce système est inspiré par des données neurophysiologiques et obéit à certaines contraintes connues sur le fonctionnement du cerveau (contrainte temporelle). Nous apportons donc une hypothèse plausible pour effectuer un traitement sur les mots pour les classer en catégories : fonction et contenu.

Par ailleurs, la méthode que nous présentons ne tient compte que de la dernière frontière d'un mot. Or, cette information ne peut exprimer la durée d'un mot. Par conséquent, c'est l'information seule de la fréquence fondamentale qui permet de catégoriser mot de fonction et de contenu. Certes la durée d'un mot influe sur la trajectoire de la F0, mais en étudiant cette trajectoire avec le TRN, la durée n'est plus nécessaire.

Ce système a été testé sur d'autres données pour l'acquisition du langage, 1) à partir de la structure prosodique pour la discrimination de classes rythmiques, 2) à partir de la structure sérielle pour la segmentation en mots, et également 3) aux structures abstraites

pour l'expérience proposée par G. Marcus et coll. (1999) (Dominey et Ramus, 2000). Ces trois expériences avaient déjà été testées chez des nouveau-nés pour montrer quelles étaient leurs capacités de traitement. Une question complémentaire est de savoir si ces traitements peuvent être réalisés par d'autres êtres vivants. Ces trois expériences ont été reconduites avec succès chez le singe (Hauser, 2002). Etant donné que l'identification lexicale peut se faire chez les nouveau-nés et par un réseau avec un traitement temporel réaliste, est-il possible de vérifier si les primates peuvent distinguer eux aussi les mots de fonction des mots de contenu, lorsqu'ils sont présentés isolément ?

V.1.2.Segmentation

Le dernier point que cette étude met en exergue est la segmentation de la parole en mot. Apparemment, l'identification lexicale ne peut se faire si les frontières des mots (ou des groupes de mots) ne sont pas indiquées. Deux alternatives se posent donc : soit nous essayons de résoudre le problème de l'amorçage syntaxique sans recourir à une segmentation ; soit nous étudions conjointement les mécanismes de segmentation et ceux de l'identification lexicale (d'autant plus que la prosodie est utile pour les deux tâches).

Pour répondre à notre première hypothèse, la présence d'un pic de F0 est fortement liée à la présence d'un mot de contenu, mais plusieurs pics peuvent appartenir au même mot. Un pic de F0 est donc le plus souvent associé à une syllabe plutôt qu'à un mot. Comme aucun indice ne marque chaque mot de contenu ou chaque mot de fonction, nous n'avons pu trouver de méthode de segmentation des mots eux-même. Comme alternative, nous proposons que cet indice soit susceptible d'indiquer la venue d'un mot de fonction en Français (Malfrère et coll., 1998). Nous avons alors étudié un prototype de F0 pour chaque voyelle, pour distinguer les voyelles en fin de groupe de mots de contenu et précédant un mot de fonction. Un score de 6 % supérieur au hasard est obtenu. Les indices prosodiques (F0 et durée des voyelles) peuvent être pris en compte pour traiter le problème de l'identification lexicale et une partie de segmentation. Notons qu'il nous faut dans ce cas, une segmentation en voyelle (Blanc et Dominey, 2004).

En outre, l'utilisation du TRN ne requiert pas d'indiquer le début d'un mot, seule la fin du mot est prise en compte par le TRN. L'information nécessaire au TRN est donc moindre par rapport aux études précédentes (Shi et coll., 1998 ; Durieux et Gillis, 2000 ; Monaghan et coll., 2003). Cette particularité est due à l'emploi du réseau TRN et d'une représentation subphonémique (une trame toutes les 5ms). Effectivement, Reali et coll. (2003) ont appuyé leur recherche sur une réseau récurrent SRN, mais les mots sont présentés un à un au réseau, ce qui implique que les deux frontières des mots soient connues.

En conclusion, il apparaît important que ce travail soit poursuivi conjointement avec une segmentation en mots pour relever de l'acquisition du langage. Cette remarque pourrait faire l'objet de recherches futures.

V.2.Extension à un nouveau corpus et une nouvelle langue

Pour le corpus LSCP en Français, les indices prosodiques situés à la fin des mots permettaient l'identification lexicale. Un des point important de notre étude est que nous avons pu valider nos hypothèses développées sur le corpus LSCP Français à partir des groupes de mots, sur le corpus MULTEXT qui comprend 10 locuteurs masculins et féminins, deux langues (Anglais et Français) et une segmentation en mots. Dans chacun de ces cas, le taux d'identification est nettement supérieur au hasard. Cependant, l'Anglais semble avoir moins souvent recours aux indices intonatifs par rapport au Français.

Le corpus MULTEXT emploie de courts passages de parole, alors que le corpus LSCP ne contenait que des phrases isolés. Hirschberg (1993) précise que tous les mots de contenu ne sont pas accentués en Anglais, particulièrement lorsqu'il s'agit de textes longs. Cette remarque pourrait expliquer une partie des différences observées entre les deux corpora.

En utilisant un algorithme qui a été démontré pour représenter les composantes macro-prosodiques de l'intonation, nous avons montré que les pics de F0 caractérisaient les mots de contenu en Français (*cf.* IV.1.3). Nous avons retrouvé une distinction entre l'Anglais et le Français, puisque les performances d'identification sont supérieures pour le Français. Ainsi, un même indice peut avoir une contribution différente suivant les langues, comme cela avait déjà été démontré par Shi et coll. (1998).

Nous avons également abordé le problème de l'identification des noms et des verbes, avec certaines des techniques élaborées avec les mots de fonction et de contenu (cf. IV.1.5). Nous prouvons ainsi que les méthodes employées peuvent se répercuter sur d'autres catégories lexicales.

Toutefois, ces résultats sont minimisés car il s'agit uniquement de parole lue [102] (Fernald et McRoberts, 1996), alors que le travail de Shi et coll. (1998) a été effectué dans le cas d'un discours non contraint.

V.3.L'hypothèse d'amorçage prosodique

V.3.1.Les conséquences pour l'acquisition de la syntaxe

Il existe deux façons de répondre aux problèmes de l'acquisition des catégories syntaxiques : la division ou le regroupement. La division peut assister des procédures de regroupement employant la sémantique (Shi et coll., 1998). Notre hypothèse part donc du principe de division. D'une seule catégorie (les mots) le traitement des indices prosodiques permet de discriminer les mots de fonction et les mots de contenu. Une fois cette distinction faite, les enfants peuvent catégoriser les mots de contenu, mais surtout cette première distinction leur permettrait d'appréhender la structure syntaxique. Dominey et coll. (2003) proposent un système artificiel qui apprend la syntaxe en employant cette

[102] Pourtant, l'application de la détection des pics de F0 au corpus développé par Plunkett indique que cette règle peut s'appliquer à la fois pour le discours lu et la parole non contrainte. Ce premier résultat très positif devrait être étayé par des études complémentaires (résultats non indiqués).

même distinction. Un tel mécanisme d'apprentissage pourrait incorporer ce système de catégorisation lexicale comme module.

Cette hypothèse d'amorçage prosodique des catégories syntaxique est soutenue par les propriétés prosodiques du parler bébé. Effectivement, celui-ci emploie des pics de F0 particulièrement prononcés (i.e. avec une amplitude de fréquence importante). Dans ce cas, les pics de F0 sont particulièrement saillants dans le signal de parole, et devraient conduire à une représentation des mots de contenu très lointaine de celle des mots de fonction. Il est vraisemblable que l'application du réseau TRN à un tel corpus montrerait que les vecteurs des mots de fonction et de contenu sont plus éloignés dans le cas du discours adressé à l'enfant, que dans le cas du discours adressé à l'adulte.

D'autre part, les indices que nous avons étudiés peuvent servir pour de nombreux autres problèmes lors de l'acquisition d'une langue. Ainsi, nous avons montré que les prototypes prosodiques pouvaient s'appliquer à la distinction des noms et des verbes en Français. Il reste dès lors possible d'appliquer nos techniques à d'autres problèmes posés par l'acquisition de la parole, comme la segmentation du signal de parole.

V.3.2.Minimalité acoustique des mots de fonction

Au cours de nos études nous avons retrouvé l'hypothèse de minimalité des mots de fonction énoncée par Morgan et coll. (1996). Effectivement la durée des mots relève de ce constat, puisque les mots de fonction ont une durée plus courte que les mots de contenu. Cet indice peut être obtenu par la multiplication de deux des indices proposés par Shi et coll. (1998) : la durée des syllabes et le nombre de syllabes. Dans notre travail, ce seul indice permet l'identification de 80 % des mots en deux catégories lexicales, que ce soit pour l'Anglais ou le Français.

En outre, le Français rend minimum les mots de fonction de manière prosodique, dans le sens où ceux-ci ne sont pas marqués par des pics de F0. Cette observation est moins souvent vérifiée en Anglais, mais reste vrai. Cette minimalité est liée à la fréquence élevée d'apparition des mots de fonction.

V.4.Perspectives

Nous pourrions reconduire plusieurs des techniques proposées pour les autres langues du corpus MULTEXT (Espagnol, Italien, Allemand). Le nombre de catégories reconnues peut également être étendues (comme par exemple pour les noms et les verbes, cf. IV.1.5.) Nous pensons cependant, que dans ce cas il faut également augmenter le nombre d'information évaluée. Par exemple, la qualité vocalique est un indice susceptible d'être pris en compte par le réseau TRN. Ce constat ouvre de nouvelles perspectives de recherches.

En outre, l'identification des mots de fonction et de contenu devrait être étudiée de manière conjointe avec des processus de segmentation du signal de parole en mots. Ainsi, il serait intéressant de tester si ces deux processus peuvent interférer l'un sur l'autre.

VI.Conclusion

Le propos de cette section était d'examiner un certain nombre d'indices prosodiques (voire acoustiques) pour l'identification de deux catégories lexicales rudimentaires : les mots de fonction et les mots de contenu. Cette discrimination fonderait le premier amorçage pour appréhender la structure syntaxique. En outre, nous proposons un modèle de réseau récurrent, avec un traitement du temps réaliste, capable d'accomplir cette distinction à partir de l'évolution de la fréquence fondamentale au cours du temps. Par rapport aux études précédentes, nous ne proposons pas l'étude de plusieurs indices, mais nous nous sommes focalisées sur quelques indices prosodiques pouvant être extrait automatiquement. Nous montrons donc le réseau TRN peut traiter l'information prosodique contenu dans une échelle temporelle locale.

La résolution de cette tâche d'identification lexicale nécessite que notre système soit sensible à des variations temporelles se produisant sur de courtes durées. Que se passe-t-il si cette sensibilité est détériorée ?

Chapitre Six Thème 4 : Simulation d'un Trouble de Traitement Temporel Auditif lors de l'Acquisition du Langage

« L'acquisition du langage est sans doute la plus grande faculté intellectuelle qui nous est demandé d'avoir. » Bloomfield, 1933.

I.Introduction

Le chapitre précédent a montré que des variations prosodiques temporelles était pertinentes pour la distinction de catégories syntaxiques, et pouvaient être extraites par le réseau TRN à partir de l'évolution temporelle de F0. Nous verrons qu'il est possible de modifier un paramètre de ce modèle pour atténuer sa sensibilité temporelle, et ainsi diminuer ces facultés de catégorisation. Dans ce contexte, le réseau TRN n'est plus à même de traiter la plus petite échelle temporelle, sur laquelle peut être définie un contour intonatif.

L'acquisition du langage n'est pas systématique : certains enfants présentent effectivement des troubles pour appréhender cette faculté intellectuelle.

Nous commencerons par une description de ces enfants présentant ces troubles (dysphasiques ou SLI) et des problèmes qu'ils rencontrent avec la parole, et plus particulièrement avec la syntaxe. Nous effectuerons ensuite une brève revue des modèles computationnels simulant des troubles du langage. Quatre expériences sont exposées, afin de simuler le comportement des enfants SLI vis à vis de la perception de la parole et des stimuli acoustiques rapides. La première montre quels paramètres peuvent être modifiés de façon à empêcher la catégorisation lexicale. Ensuite, nous montrerons que le modèle discrimine toujours des séquences ayant des éléments de longues durées. Enfin, nous répliquerons deux expériences de distinction de stimuli auditifs, pratiquées avec les enfants SLI.

Notre objectif est donc de montrer qu'il est possible d'implémenter une architecture de réseau récurrent temporel, qui rend compte à la fois de troubles pour la catégorisation lexicale, ainsi que pour le traitement des stimuli brefs, constitués de tons purs, tout en conservant ces capacités de traitement intacte pour les échelles temporelles supérieures.

II.Troubles du langage : le cas des enfants SLI

Les enfants sont diagnostiqués dysphasiques (ou SLI : Specific Language Impairment) [103] lorsque l'acquisition du langage s'effectue plus lentement que la normale sans raison évidente. Par définition, ces enfants ne souffrent pas d'un trouble permanent de l'audition. Ils sont atteints de troubles spécifiques du langage c'est-à-dire de limitations significatives dans le développement et la maîtrise de la parole, ainsi que de la lecture. Pour autant, aucune condition manifeste, comme un QI relativement bas, ou une audition endommagée, ne peut être décelée chez ces enfants. Aux Etats-Unis, entre 3 et 7 % des enfants seraient touchés par ces troubles.

La question des troubles des enfants SLI se fonde sur deux points : les difficultés qu'ils ont avec le langage et les hypothèses sur leur traitement du langage. Une de ces hypothèses privilégie un trouble du traitement auditif rapide. Cette théorie sera examinée, avant d'être discutée.

II.1.Leurs difficultés avec le langage

Les enfants SLI exhibent des difficultés à la fois pour le traitement du langage oral, mais aussi pour la lecture. Leurs difficultés liées au traitement de la grammaire seront examinées plus attentivement.

II.1.1.Trouble de la parole

Les enfants SLI sont en retard pour l'acquisition de leurs premiers mots (Leonard, 1998).

[103] Nous emploierons le terme d'enfants SLI plutôt que dysphasiques, car notre travail s'inspire en majorité de la littérature Anglo-saxonne.

Par rapport aux enfant normaux, ils produisent plus souvent des erreurs et des sons inhabituels (Leonard, 1985) : ils remplacent le son /v/ par /d/ et des nasales sont ajoutées en position initiale ou finale. Quelquefois des problèmes pour trouver les mots sont mentionnés. Leonard (1998) dresse un tableau complet des troubles des enfants SLI.

II.1.2.Trouble de la lecture

Quelques recherches partent du principe que les difficultés spécifiques à la lecture (Specific Reading Disability SRD) sont un type de troubles du langage, tandis que d'autres suggèrent que les enfants dysphasiques et dyslexiques (respectivement, SLI et SRD pour la communauté anglophone) sont différentes manifestations du même trouble. Ainsi, la plupart des recherches suggèrent de former trois groupes :

1. Les enfants SLI ;
2. Les enfants présentant seulement des difficultés pour la lecture (SRD) ;
3. Et la majorité des cas regroupant les enfants présentant les deux troubles simultanément, sous le sigle LLI (Language based Learning Impairment). Effectivement 80 % des enfants SLI vont présenter des troubles pour la lecture.

Toutefois, notre attention se portant uniquement sur les bases du développement de la grammaire, nous ferons uniquement référence aux enfants SLI.

II.1.3.Déficit grammatical

Les troubles touchant la grammaire s'articulent aussi bien autour de l'acquisition des structures grammaticales, que des catégories fonctionnelles. Ces troubles ont principalement été étudiés chez les enfants SLI anglophones, cependant certaines difficultés se retrouvent chez les enfants s'exprimant dans d'autres langues.

II.1.3.1.Structure grammaticale

Les enfants SLI n'utilisent pas certaines des constructions observées dans le discours des enfants normaux. Ils auraient un ensemble plus restreint de règles syntaxiques (Lee, 1966). Bishop (1979) a fait passé des tests de vocabulaire et de compréhension grammaticale à des enfants présentant des troubles de la production et à des enfants contrôles choisis d'après un test d'intelligence non verbale. Les enfants SLI ont montré des difficultés de compréhension uniquement dans le cas des phrases passives. Les phases actives ou les groupes prépositionnels dont les mots sont ordonnés de façon canonique sont aussi bien compris par les deux groupes d'enfants. Selon Van der Lely (1996), les enfants SLI ont plus de difficultés pour déterminer les sujets et les agents à partir de la structure syntaxique seule. Gopnik (1990b) a suspecté que les enfants SLI n'avait pas de connaissance grammaticale du rôle des morphèmes. Effectivement les enfants SLI qui ont été étudiés ont montré des erreurs sur les inflexions et les mots de fonction, qui désignent le temps, l'indéfini, la personne, le nombre et le genre (Gopnik, 1990b).

II.1.3.2.Catégorie fonctionnelle

Les individus SLI ont des problèmes plus particulierement pour l'acquisition des catégories fonctionnelles (Eyer et Leonard, 1995 ; Guilfoyle, Allen et Moss, 1991 ; Leonard, 1995). Les morphèmes grammaticaux qui posent le plus de problèmes pour les enfants SLI parlant Anglais sont tous associés à des catégories fonctionnelles. Le trouble le plus commun est l'omission des inflexions et des mots de fonction. Le plus grand problème pour ces enfants réside dans leur capacité limitée pour retenir et traiter des éléments grammaticaux subtils comme l'inflexion 'ed' dans « the horse jumped over the fence » ou des mots de fonction comme 'is' dans « the horse is running », et non pas dans leur capacité de compréhension des principes grammaticaux (Leonard, 1998).

Avec l'analyse de la structure syntaxique proposée par Morehead et Ingram (1973), il est possible de déterminer le degré d'utilisation de certains mots de fonction. Ainsi l'auxiliaire *do* est utilisé moins fréquemment par les enfants SLI, que par les contrôles MLU [104] . A partir de la même base de données Ingram (1974) a montré que les enfants SLI utilisaient l'auxiliaire *be* avec un pourcentage inférieur dans les contextes obligatoires, par rapport aux enfants MLU. Il s'agissait la plupart du temps d'omissions.

Les mots de fonction ne sont pas totalement absents du discours des enfants SLI. Toutes les différences entre les enfants SLI et les contrôles MLU proviennent d'un degré inférieur d'utilisation des catégories fonctionnelles dans les contextes obligatoires. Fletcher (1983) pense que les omissions des mots de fonction seraient dues à leurs caractéristiques phonétiques, en particulier leur brièveté qui tend à les rendre moins saillants dans le signal de parole. La plupart des morphèmes de classe fermée omis par les enfants SLI avaient aussi une durée très courte [105] (Leonard, 1989).

Les enfants SLI sont lents pour maîtriser certaines formes pronominales. Rispoli (1994) a proposé qu'au cours du développement normal de la parole, les enfants ne remplacent pas un pronom par un autre dont le matériel phonétique diffère. A l'inverse, un pronom peut être substitué par un autre qui partage les mêmes propriétés phonologiques. Ogiela (1995) a prouvé ce point de vue avec les enfants SLI. Effectivement, ces enfants produisent plus souvent *her* dans un contexte demandant *she*, que *him* for *he*.

Les mots de fonction, qui sont critiques pour le processus de traitement grammatical subissent des traitements différents chez les enfants SLI, ayant des troubles grammaticaux. En effet, une étude en Potentiel Evoqué a révélé qu'ils produisent une négativité latéralisée à droite, alors qu'elle apparaît à gauche chez les enfants normaux (Ullman et Pierpont, In Press).

II.1.3.3.Le cas des langues autres que l'anglais.

La sévérité des troubles chez les enfants SLI dépend de la nature de la langue apprise.

[104] Les enfants MLU (Mean length utterance) produisent des phrases ayant le même nombre de syllabes que les phrases énoncées par les enfants SLI. Les contrôles sont donc le plus souvent plus jeunes que les enfants SLI.

[105] La troisième personne du singulier –*s*, des articles, des formes auxiliaires *be*, de la marque de l'infinitif *to* et la préposition *that*.

Ainsi, la grammaire morphologique est fortement perturbée dans les langues très infléchies comme l'Hébreu, l'Italien, ou le Serbo-croate. Dans le cas de systèmes morphologiques riches comme l'Allemand, de nombreux indices permettent de prédire le genre ou le cas du mot suivant un article. Deux langues opposées par leur morphologie (l'Anglais et l'Allemand) ont été examinées. Il s'est avéré qu'une réduction globale des capacités de traitement affectait certains aspects de la parole plus que d'autres suivant la langue (Kilborn, 1991).

En Français, les mots de fonction constituent moins de 10 % des types lexicaux. Les enfants SLI francophones produisent les mêmes erreurs que les enfants anglophones. La plus grande difficulté semble portée par le pronom *il.* Il ne s'agit vraisemblablement pas d'un problème de cas, puisqu'ils ne rencontrent pas de difficultés avec le pronom *elle*. Cependant, ils éprouvent moins de difficultés avec les articles en Français (Lenormand, Leonard et McGregor, 1993), qui sont plus saillants que les articles en Italien. En Français, la durée des articles et des syllabes non accentuées diffèrent relativement peu des syllabes accentuées non finales (Fant, Kruckenberg et Nord, 1991). En outre, beaucoup de mots français commencent avec une syllabe faible suivie d'une forte, ce qui les prépare pour l'utilisation des articles. En revanche, il reste possible que les articles soit appris par cœur avec le nom qui les suit.

II.2.Quels sont les points critiques de la discussion ?

La discussion des origines des troubles des enfants SLI est centrée sur la nature du problème : soit un problème d'encodage ou de traitement de la parole, soit un déficit général (i.e. observé pour plusieurs modalités) ou spécifique au langage, soit d'autres facteurs extérieurs au langage comme la cognition sociale.

II.2.1.Problème d'encodage ou de traitement ?

Le débat sur les enfants SLI repose essentiellement sur l'origine de leurs troubles. Une fois cette origine élucidée, des solutions pourront être proposées pour résoudre leurs problèmes.

Deux hypothèses [106] ont été établies :

1. Un déficit du traitement de bas niveau effectué par les organes sensoriels (Tallal et Miller, 1993) ;
2. Un trouble des fonctions cognitives de haut niveau.

La question est donc de déterminer si les troubles du traitement auditif des informations rapides pourraient être dus à des anomalies lors de l'encodage neurophysiologique des différences acoustiques de la parole. Celles-ci auraient lieu après le système périphérique sensoriel et avant la perception consciente.

[106] Des hypothèses alternatives ont aussi été proposées. Par exemple, les enfants SLI pourraient être conscients des règles gouvernant le langage, mais auraient des capacités de traitement limité (Leonard, 1998). En résumé, les connaissances gérées automatiquement par les enfants normaux demanderaient plus d'effort aux enfants SLI

Les troubles des enfants SLI pourraient être la conséquence d'un mécanisme défectueux, spécifique d'une espèce et spécialisé pour l'apprentissage du langage. Le cas des enfants SLI serait une opportunité pour comprendre les particularités du système d'apprentissage de la parole et pour savoir ce qui distingue ce système des autres fonctions cognitives.

II.2.2.Un déficit général ou spécifique du langage ?

Tallal et ses collègues postulent que les enfants SLI sont caractérisés par des troubles du traitement auditif pour les événements se déroulant rapidement. La théorie du déficit temporel ne serait pas restreinte à la modalité auditive. Ceci pourrait même se traduire par un dysfonctionnement global, pour la perception et la production de toutes les informations sensori-motrices brèves (Habib, 2002). Ainsi, de nombreuses expériences ont démontré des déficits observés en modalités visuelles, particulièrement pour les enfants dyslexiques et mêmes pour les modalités tactiles (Tallal, Stark et Mellits, 1985). Ainsi, les troubles de la morphologie grammaticale résulteraient d'une limitation générale pour les traitements perceptuels et cognitifs.

Les troubles observés pourraient refléter un déficit spécifique du langage et non de l'audition (Mody, 1993). Dans ce cas, il s'agit de savoir si les difficultés proviennent de la complexité du stimulus ou si cela est dû à la nature spécifique du signal de parole. Les performances des enfants SLI devraient être examinées face à des stimuli verbaux et non-verbaux de même complexité.

En résumé, la question de connaître si les troubles des enfants SLI ont pour origine un déficit spécifique du langage, ou des capacités de traitement auditif limité pour tous les sons, est toujours au cœur du débat. La notion de déficit perceptuel comme source des troubles du langage est apparue avec Lowe et Campbell (1965).

II.3.L'origine des troubles des enfants SLI expliqué par un déficit du traitement auditif temporel

Les troubles des enfants SLI peuvent être la conséquence d'un déficit du traitement rapide des évènements en particulier pour la modalité auditive. Quatre points vont être examinés pour étayer cette hypothèse. Tout d'abord, les troubles liés à ce déficit seront résumés, avant de présenter ceux concernant la grammaire. Ensuite, les expériences prouvant ce déficit dans le traitement des évènements rapides seront décrites. Enfin, un point sur les techniques de rééducation inspirées par cette théorie conclura la présentation de l'hypothèse de trouble du traitement rapide.

II.3.1.Quels sont les troubles observés ?

Les enfants SLI montrent plus de difficultés lorsque le matériel verbal leur est présenté rapidement, ainsi que pour le traitement de quelques phonèmes et syllabes. En outre, ces déficits gardent encore une certaine influence à l'âge adulte.

II.3.1.1.Vitesse de présentation

Les mots infléchis présentés dans des contextes variés (Haynes, 1982) et dans des phrases prononcées rapidement (Elis Weismer et Hesketh, 1993) ont un effet plus perturbant pour les enfants SLI que pour les enfants normaux. L'apprentissage des mots nouveaux est facilité dans des phrases énoncées lentement (Elis Weismer et Hesketh 1993, 1996).

II.3.1.2.Traitement des phonèmes et des syllabes

Les enfants SLI ne peuvent couramment pas identifier des stimuli rapides à l'intérieur de la parole. Des durées inférieures à quelques dizaines de millisecondes ne peuvent être traitées par ces enfants. Or, cette durée correspond justement à de nombreux contrastes phonétiques [107] . Les enfants SLI peuvent identifier les deux phonèmes [tʃ] et [ʃ] lorsqu'ils sont présentés isolément (Leonard, 1998). Il ne s'agit donc pas d'un problème de discrimination des phonèmes.

Tallal (2000) exprime la même hypothèse pour une fenêtre temporelle plus large (de l'ordre de quelques centaines de millisecondes). Dans ce cas, la perception des syllabes est perturbée. Ces difficultés de traitement pourraient être issues d'une capacité de segmentation détériorée. La plus petite unité de traitement de la parole serait de l'ordre de la syllabe et non du phonème (Bishop, 1997). Leonard (1998) stipule que les enfants SLI rencontrent des difficultés avec des inflexions grammaticales (le suffixe du passé *–ed* en anglais) et aussi des mots de fonction (l'auxiliaire *is*), mais ont la capacité de comprendre les principes grammaticaux. Les mots et les morphèmes fonctionnels sont moins saillants perceptuel-lement. Dans ce cas, un système de traitement auditif aurait plus de difficulté pour les percevoir. Du coup, les enfants recrutent plus de ressources pour les percevoir et les produire.

Dans ces deux cas, l'hypothèse d'un déficit dans le traitement auditif temporel est retenue au détriment de l'explication d'un déficit de traitement cognitif de haut niveau. De nombreux études ont d'ores et déjà validé ces hypothèses pour au moins un sous-groupe d'enfants SLI. Effectivement, elles montrent que ces enfants ont des difficultés pour discriminer des stimuli acoustiques (verbaux ou non), en particulier lorsque la différence réside dans des durées inférieures à la centaine de millisecondes. Tallal et coll. (1985) concluent que les caractéristiques temporelles des stimuli acoustiques sont critiques pour les enfants atteints de troubles du langage. Quand les stimuli sont brefs ou présentés rapidement, les enfants SLI ont des difficultés pour les différencier, alors qu'ils n'en ont pas pour les distinguer lorsqu'ils sont allongés ou présentés plus lentement. Ce déficit est désigné par les termes d'un trouble du traitement auditif temporel (Benasich et Tallal, 2002).

En outre, le lien entre les facultés de traitement des événements rapides et les facultés pour le langage (richesse du vocabulaire, structure des phrases plus complexes,

[107] La distinction des phonèmes se fait par des indices qui varient très rapidement dans le temps (moins de quelques centaines de millisecondes).

plus de mots irréguliers employés) a été établi également pour des enfants normaux, qui effectuait une tâche de détection d'intervalle (« gap detection ») avec des scores supérieurs à la médiane (Trehub et Henderson, 1996).

II.3.1.3.Un déficit durable

Les connaissances acquises pour la maîtrise du langage restent fragiles chez les enfants SLI. Effectivement, les adultes ayant eu des troubles du langage dans leur jeunesse ont de faibles performances, lors de tests contenant beaucoup d'informations nouvelles (Tomblin, 1994 ; Bishop et coll., in press). Ainsi, un déficit apparaissant tôt dans le développement peut avoir un impact important sur l'organisation cognitive, mais si celui-ci est résolu peu après son apparition, il disparaîtra totalement.

II.3.2.Lien entre la grammaire et le déficit de traitement rapide

Les capacités de discrimination des enfants SLI ont été étudiées sur des formes grammaticales de courtes durées (*-s, -ed, he, she, they …)* et de durées plus longues (*him, her, more …)*. Comparativement aux enfants MLU, les enfants SLI montrent plus de difficultés avec les formes grammaticales de courtes durées. La seule autre forme à poser problème aux enfants SLI est le passif *the clown is being pushed* (Fellbaum, Miller, Curtiss et Tallal, 1995).

Bien que le groupe des enfants SLI soit hétérogène, la majorité des erreurs sont en lien avec l'aspect formel du langage, en particulier la syntaxe et la phonologie. Les enfants ne pourraient analyser la parole pour des niveaux inférieurs à la syllabe. En conséquence, les enfants SLI sauraient que *cats et dogs* font référence à chaque fois à plusieurs chats ou chiens, sans savoir qu'il existe une règle de formation du pluriel en ajoutant le morphème *–s* (Gopnik, 1990a). Fletcher (1983) postule que les enfants SLI ont des problèmes avec l'utilisation des auxiliaires, parce que ceux-ci n'apparaissent pas distinctement dans le signal. D'autre part, la brièveté des inflexions en Anglais ne facilite pas l'acquisition de la morphologie grammaticale. Les morphèmes fonctionnels qui ne sont pas utilisés par les enfants SLI sont tous de courtes durées (Leonard, 1989).

II.3.3.Les expériences pour tester l'hypothèse de déficit temporel

Afin d'appuyer la théorie du déficit de traitement rapide, de nombreuses expériences de perception ont été menées avec des tests utilisant aussi bien des stimuli acoustiques comme des tons purs, que de la parole.

II.3.3.1.Tests avec du matériel non verbal

Des stimuli non-linguistiques ont été employés pour mettre à l'épreuve la théorie du déficit de traitement rapide. Nous allons établir un panel représentatif des divers paradigmes utilisés :

- Identification de l'ordre dans des paires de tons purs avec des fréquences distinctes (Rapid Perception Test : Tallal et Piercy, 1973a) ;

- Détection d'un bref silence dans un son continu (Gap detection). Dans ce cas, les enfants SLI ne détectent pas le silence si celui-ci a une trop courte durée ;
- Fusion de deux pulsations en seul événement (auditory fusion), testée seulement sur des dyslexiques ;
- Repérage dans l'espace d'un clic sonore (auditory tracking, enfants SLI) ;
- Tons purs avec des différences de modulation et d'amplitude ;
- Apprentissage de motifs acoustiques (Tomblin et Quinn, 1983) ;
- Discrimination de tons purs ayant ou non la même fréquence (Same-Different Task ; Tallal et Piercy, 1973b). Les variables sont les durées de l'intervalle entre les deux tons (ISIs) et la durée de chacun d'eux. Les enfants SLI ont des difficultés pour distinguer les tons pour les durées ISIs les plus courtes [108] ;
- Détection d'un ton bref, suivi d'un masque de bruit (Backward et Forward Masking, Wright et coll., 1997, enfants SLI). Cette expérience montre des différences entre les enfants contrôles et SLI uniquement pour une condition (Backward Masking). Il ne peut donc s'agir de déficit attentionnel, mais bien un problème lié au traitement auditif [109] .

II.3.3.2.Tâche avec du matériel linguistique

Quelques expériences ont été conduites à partir de stimuli verbaux. Ainsi, la discrimination de la paire [ba] – [da] était une des six tâches qui distinguaient les enfants SLI, des enfants se développant normalement, par l'intermédiaire d'une fonction d'analyse discriminante (Tallal, Stark et Mellits, 1985). La façon dont sont synthétisées [110] les syllabes influe sur leur perception par les enfants SLI (Henderson, 1978). La portion acoustique qui distingue les deux stimuli était très courte par rapport aux restes de la syllabe (Leonard, McGregor et Allen 1992). Les enfants SLI ont des difficultés de discrimination pour la paire /ba-da/ pour des transitions formantiques courtes (43 ms) et non pour des durées plus longues (95 ms). La perception serait améliorée par l'accroissement du temps de traitement autorisé par l'allongement du stimulus (Tallal et Piercy, 1975).

II.3.4.Rééducation

Les enfants SLI améliorent leur performance avec de l'entraînement dans les tâches de perception auditive rapide (Tomblin et Quinn, 1983 ; Tallal et coll., 1981). En outre, ces améliorations ont été constatées pour des stimuli de parole modifiée, mais aussi pour de la parole naturelle. L'apprentissage dépend des sujets, mais certains atteignent l'ordre d'une dizaine de millisecondes pour les transitions formantiques, seuil observé chez les sujets sains (Tallal et coll., 1996). Un accroissement significatif des capacités linguistiques (environ 2 années) a pu être observé pour une période d'apprentissage de 4 semaines (Tallal et coll., 1996). Ces progrès indiqueraient que les enfants SLI ne souffrent pas de troubles biologiques ou moléculaires irrémédiables (Merzenich et coll., 1996). L'étude des

108 Ce phénomène est encore amplifié lorsque d'autres éléments acoustiques sont pris en compte dans la tâche (succession de 3 tons purs).

109 Un sujet a réalisé les pires performances dans cette condition, et c'est aussi lui qui présentait de sévères difficultés pour la lecture des mots et non-mots et pour épeler les mots.

110 Les deux variables en jeu étaient le voice onset time et le décroissement du premier formant.

Potentiels Evoqués a montré que les ondes MMN (MisMatched Negativity) changeaient de forme après l'apprentissage (Leonard, 1998).

Les enfants SLI ont besoin d'un nombre plus élevé d'essais, pour atteindre le seuil des enfants du même âge, dans des tâches de détection de silence. Les résultats de ces expériences montrent que les enfants SLI perçoivent les aspects temporels des stimuli acoustiques aussi bien que les enfants se développant normalement. Ainsi, les enfants SLI ne sont pas totalement incapables de traiter les événements rapides, mais ont besoin d'une période d'apprentissage plus longue.

II.4.Critique de la théorie de déficit du traitement rapide

Les preuves d'un trouble général de la perception auditive chez les enfants SRD et SLI ont été discutées de nombreuses fois. Certes, la parole est composée de nombreux détails acoustiques constitués d'indices de courtes durées. Il serait alors raisonnable de penser que l'on puisse obtenir une relation entre leurs performances pour des tâches phonologiques et leurs capacités pour la parole. Mais la parole comporte aussi des indices qui ne sont pas encodés dans des courtes durées. La plupart du temps des indices additionnels sont présents, lorsque la distinction repose sur des indices brefs (Leonard, 1998).

De plus, la difficulté de compréhension des contrastes grammaticaux apparaît également chez les enfants profondément sourds (Bishop et coll., 1999). Ludlow et coll. (1983) ont trouvé des déficits chez les enfants souffrant de problèmes attentionnels, lors d'une tâche de jugement d'ordre temporel, et Howell et coll. (1999) ont exhibé des déficits dans les tâches de détection après un masque, pour les personnes qui bégaient. Il se peut toutefois que tous ces groupes distincts souffrent d'un déficit commun pour les tâches nécessitant un traitement auditif rapide (Leonard, 1998).

En outre, la difficulté observée pour les évènements rapides est mise en doute chez les enfants SLI. Des études ont montré des troubles pour les processus auditifs en général, sans dénoter de difficultés plus particulières pour les événements rapides (Rosen et Manganari, 2001 ; Bishop et coll., 1999). Ces deux études ne retrouvent également pas les résultats avancés par Wright et coll. (1997). La fonction d'analyse discriminante classant les enfants SLI (Tallal et coll., 1985) n'a pas été validée sur de nouvelles données.

La théorie du déficit de traitement rapide postule que le déficit est général et cette même hypothèse n'explique donc pas la nature apparemment sélective des troubles non-linguistiques. En dernier lieu, tous les types de troubles pourrait être expliqués par des capacités de traitement limitées ou par un ralentissement général. En revanche, il n'est pas clair que tous les enfants SLI souffrent de ces problèmes. Ainsi, les données expérimentales suggèrent que le déficit de traitement est fortement associé aux troubles du langage, mais qu'il n'apparaît pas nécessairement en même temps que ces troubles (Ullman et Pierpont, In Press ; Bishop et coll., 1999). Ainsi, certains enfants SLI ne présentent aucune difficulté pour le traitement des évènements rapides. En outre, des enfants avec des compétences normales pour le langage montrent des difficultés

équivalentes aux enfants SLI, lors de tests de traitement rapide (Bishop et coll., 1999).

Lors des tests réalisés pour appuyer l'hypothèse d'un déficit du traitement rapide, un certain de nombre de paramètres ont été laissés libres. Chacun des paragraphes suivants retracent chaque point critiquable de ces études :

- **L'attention et la mémoire.** Le traitement auditif rapide fait appel des capacités non perceptives telles que la mémoire, l'attention et également la classification des stimuli. Ces capacités sont souvent sous-développées chez les populations étudiées. Quatre des six variables temporelles distinguant les enfants SLI des enfants normaux (Tallal et coll., 1985) tiennent compte des capacités de la mémoire de travail.
- **La charge cognitive induite par les expériences.** Il est possible que les déficits observés lors des traitements rapides, soit dus à l'augmentation de la pression induite par la tâche.
- **Sensibilité de la tâche contrôle.** Les difficultés de la tâche contrôle et de la tâche de test doivent être équivalente. Pour la tâche présentée dans Tallal et Piercy (1973a), tous les sujets atteignent le seuil maximal de 100 %, lorsque l'intervalle inter-stimulus est élevé. Cela révèle un effet plafond pour ces performances. La plupart des tâches, créées pour tester les capacités de traitement rapide ont une variation moins importante des performances, pour les tâches contrôles. La difficulté de ces tâches devrait être augmentée lors des essais à vitesse plus lente.
- **Le groupe contrôle.** La plupart du temps, les personnes entrant dans le groupe contrôle sont coutumières de ce genre de tâche. Etant donnée qu'un entraînement permet d'améliorer les performances, les scores de ce groupe seraient plus élevés que la moyenne. En conséquence, ce ne sont pas les enfants SLI qui ont performances faibles, mais les enfants du groupe contrôle qui ont des performances remarquables.
- **L'âge et les capacités d'apprentissage.** Le seuil de détection des silences prédit le développement du langage chez les bébés normaux (Trehub et Henderson, 1996). Ces seuils de détection sont également élevés chez les enfants SLI et SRD, bien qu'il n'y ait pas de distinction entre les adultes normaux et les adultes dyslexiques. Il est possible que les mesures du traitement auditif rapide indiquent une attention insuffisante, un apprentissage plus lent, plutôt qu'une limitation fondamentale de la perception. Ceci est encore compliqué par le fait que le traitement auditif change avec l'âge. Testés pour la première fois à 6 ans, 19 enfants avec des difficultés d'apprentissage parmi 29 n'identifient pas de sons synthétiques, alors qu'au deuxième test 24 d'entre eux atteignent le critère de distinction. Ces performances étaient atteintes alors que 23 des enfants avait des scores psychométriques justifiant de leur difficultés d'apprentissage (Clark et coll., 2000).

II.4.1.Les enfants SLI : une condition hétérogène

D'un point de vue général, les populations SLI testées sont relativement peu décrites. Il s'agit souvent d'enfants présentant des troubles de dysphasie, dyslexie, ou des difficultés

d'apprentissage. Les groupes cliniques d'enfants SLI sont assez mal définis. Entre autres, certains enfants SLI ont des difficultés pour la morphologie régulière ; un sous-groupe d'enfants SLI aurait des troubles avec les structures syntaxiques, mais pas pour la morphologie. Les scores pour les tâches de traitement rapide sont souvent répartis dans une plage plus grande que celle observée chez les contrôles.

II.4.2.L'hypothèse du déficit du système procédural

Aucune des théories précédentes expliquant les troubles des enfants SLI ne peut facilement décrire la variété des troubles affectant les fonctions linguistiques et non-linguistiques, et ce même à l'intérieur des sous-groupes d'enfants SLI. En outre, peu de ces théories ont tenté de relier ces troubles cognitifs avec le système nerveux. Les troubles du langage observés chez les enfants SLI seraient la conséquence de nombreux facteurs qui agissent en synergie. Ainsi, un déficit de traitement rapide ne pourrait avoir des conséquences que pour les enfants présentant d'autres troubles (risques génétiques, etc.). Dans ce contexte, il a été proposé que les enfants SLI aient des altérations du système nerveux dédié à la mémoire procédurale (Ullman et Pierpont, In Press)

Le système procédural est composé d'un réseau de plusieurs structures cérébrales interconnectées situé dans les circuits entre le cortex frontal et les glanglions de la base. Ce réseau permet de contrôler et d'apprendre des fonctions cognitives et motrices habituelles, telles que la marche, le vélo, les jeux, ou l'écriture.

Les enfants SLI exhibent des troubles dans toutes les tâches dépendant du système procédural, ce qui soutend qu'ils possèdent un système procédural déficient. En outre, les troubles diffèrent suivant les mécanismes cérébraux atteints. Sous cette hypothèse, certains enfants SLI peuvent présenter un déficit du traitement rapide alors que d'autres n'en présentent pas. En outre, les particularités anatomiques des structures frontales et des ganglions de la base qui ont souvent été remarquées chez les enfants SLI corroborent l'hypothèse du déficit procédural (Ullman et Pierpont, In Press).

II.5.Conclusion

L'origine des déficits du traitement dans les populations SLI est toujours un point de discussion. En résumé, cela pourrait provenir de déficits dans les traitements sensoriels de base (Tallal et coll., 1996) ou de processus cognitifs de plus haut niveau. La première hypothèse postule que des anomalies dans le codage neurophysiologique des différences acoustiques de la parole sont responsables de ce déficit. Cet affaiblissement du traitement temporel auditif a été montré par plusieurs expériences auditives comprenant des stimuli non-verbaux. Quand ceux-ci sont brefs ou rapides, les enfants SLI ont des difficultés pour les distinguer, bien qu'ils n'aient aucune difficulté pour différencier les mêmes stimuli quand ils sont rallongés ou présentés à une vitesse plus lente. Au moins un sous-groupe d'enfants SLI exhibe un déficit pour le traitement des événements auditifs rapides en même temps qu'il présente des difficultés lors de la manipulation des mots et des phonèmes fonctionnels (Benasich et Tallal, 2002). Les disparités observées chez les enfants SLI proviendraient partiellement des portions atteintes du système procédural

(Ullman et Pierpont, In Press).

Cette section a présenté les troubles du langage observés chez les enfants SLI. Outre les enfants SLI, de nombreux cas de dysfonctionnement du langage ont été remarqués, quels modèles du traitement du langage permettent de simuler ces troubles ?

III.Etat de l'art : simulation des dysfonctionnements du langage

Trois types de modèles peuvent être proposés pour simuler le dysfonctionnement de la perception de la parole : l'être humain soumis à de la parole dégradée, l'animal lors de la perception de stimuli non-verbaux et des modèles informatiques de traitement du langage.

III.1.Modèle adulte

Le profil syntaxique observé chez les enfants SLI anglophones peut être répliqué chez des adultes sans trouble apparent. La parole doit être comprise, soit dans des conditions d'écoute difficiles (Kilborn, 1991), soit dans des conditions où les ressources cognitives sont partagées avec d'autres tâches (Blackwell et Bates, 1995). Dobie et Berlin (1979) ont montré que les phrases dont l'intensité a été réduite de 20 dB contiennent une information acoustique beaucoup plus réduite pour les mots de fonction et les inflexions [111] .

III.2.Modèle animal

Les rats ayant des lésions microgyriques ont des difficultés pour le traitement des informations acoustiques rapides. L'étude de certaines malformations bien définies du cortex chez les rongeurs permet de mieux comprendre les déficits du traitement des sons non linguistiques. Les animaux qui ont des mycrogéries [112] ne peuvent distinguer des sons quand ils sont séparés par des intervalles de temps très courts (Galaburda, 1994). L'induction de malformations corticales entraîne des modifications dans le Thalamus qui, elles-même sont associées à des déficits de la discrimination des sons. Ces facultés de traitement sont en outre liées à la complexité du stimulus (Clark et coll., 2000). Toutefois, ces résultats ne sont pas mis en relation avec l'âge ou l'entraînement lors des tâches de discrimination de stimuli rapide (Clark, Rosen, Tallal et Fitch, 2000).

III.3.Modèle informatique

[111] Par exemple, la version atténuée de la phrase *Where are Jack's gloves to be placed?* devient *Where Jack glove to be place*?

[112] Troubles de l'organisation des couches du cerveau.

Des erreurs observées chez tous les enfants, comme des phénomènes de surgénéralisation lors de l'apprentissage des formes au passé des verbes en Anglais, ont été simulées par un réseau de neurones (Plunkett et Marchman, 1993). Des déficits ont également été montrés lorsque des lésions sont simulées par un changement de paramètre du modèle. Ainsi, les réseaux connexionnistes pouvaient avoir une grammaire endommagée soit en dégradant les entrées (Hoeffner et McClelland, 1993), soit en réduisant la proportion de connexions (Marchman, 1993). Ce dernier modèle a été proposé pour rendre compte des performances observées chez des aphasiques. Haarman, Just et Carpenter (1997) proposent également un modèle connexionniste, qui compare l'aphasie à un manque de ressource pour assigner les rôles thématiques à partir de la structure syntaxique. Ce déficit est souvent expliqué par une réduction des ressources de la mémoire de travail, simulée ici par une réduction du nombre de connexions.

Le nombre de modèles informatiques proposés pour rendre compte des déficits observés chez les enfants SLI est plus restreint. L'hypothèse de déficit de traitement rapide est souvent représentée comme un allongement de la fenêtre d'analyse des récepteurs auditifs, ce qui se traduit par un manque de précision temporelle. Hartley et Moore (2001) démontrent que ce seul facteur ne peut être la cause de cet handicap. Il faut tenir compte du processus de traitement, de l'attention et de la motivation. Ils retrouvent alors les résultats pour les tâches de masquage antérograde (Backward Masking). D.V.M Bishop et K. Plunkett projette de simuler l'apprentissage des inflexions grammaticales à partir d'un réseau, dont les entrées sont dégradées [113] .

IV.Matériel et méthode

IV.1.Corpus

L'expérience de catégorisation lexicale est menée sur le corpus LSCP Français, avec les groupes de mots de même nature lexicale (fonction ou contenu, cf. chapitre 5, section III.1.1).

IV.2.Représentation des données

Pour l'expérience de catégorisation lexicale, la fréquence fondamentale est extraite par intervalles de 10 ms, en utilisant le logiciel BLISS, de John Mertus. Ces valeurs numériques sont ensuite transformées à l'aide d'une courbe de Gauss (cf. chapitre 3, section III.3).

Pour les tâches avec des séquences abstraites, un évènement est associé à un

[113] THE NATURE AND CAUSES OF LANGUAGE IMPAIRMENTS IN CHILDREN A programme of research funded by the Wellcome Trust : Principal Investigator : Dorothy V. M. Bishop.

neurone de la couche d'entrée.

Dans le cas de la réplication de la tâche de Tallal et Piercy (1973a), les tons purs sont représentés à l'aide d'une courbe de Gauss, comme pour l'expérience de catégorisation lexicale.

Pour la réplication de l'expérience de Wright, nous utiliserons d'abord une représentation abstraite des données, où le bruit est représentée par une activation uniforme des neurones d'entrées du réseau, et l'intensité est considérée sur une échelle linéaire. Ensuite, nous aurons recours au cochléogramme pour encoder les stimuli auditifs directement, sans passer par un codage manuel. Le cochléogramme est obtenu à partir de Matlab et de l'implémentation du modèle de cochlée développé par Meddis dans la boîte à outils réalisée par M. Slaney (1998). Les sons ont été générés à partir du logiciel Matlab, et en suivant les indications données dans Wright et coll. (1997).

IV.3.Méthodes de traitement

Ce chapitre ne fait pas appel à de méthodes particulières utilisées avec le TRN. Effectivement, nous interviendrons au niveau des constantes de temps, définissant les réponses dans le temps des unités du réseau. Plus une constante de temps est élevée, plus la réponse est lente à atteindre le niveau du signal d'entrée (cf. Figure 1.2).

Suivant la couche où se situent les unités « ralenties », l'effet sur le traitement effectué par le réseau devrait être différent. Si elles appartiennent à la couche $State_D$ les évènements éloignés dans le temps ne doivent plus influencer les évènements courants. Dans le cas de constantes de temps infinies, cela revient à supprimer les connexions récurrentes, reliant les couches State et $State_D$. Dans le cas où les éléments de la couche State sont modifiés, les évènements rapides ne devraient plus être perçus, mais le traitement global de l'information lente reste possible. Ainsi, une des extrémités du Continuum Temporel traité par le réseau est endommagée, alors que les autres domaines temporels restent intacts.

V.Expérimentation : Simulation d'un déficit temporel

Dans le chapitre précédent, nous avons montré que des indices prosodiques peuvent permettre la distinction entre les mots de fonction et les mots de contenu. En outre, ces indices peuvent être traités par un modèle de réseau récurrent temporel pour opérer cette identification, justifiant ainsi le dernier point de notre hypothèse de Continuum Temporel. Nous allons voir maintenant si il peut exister une implémentation informatique, sous forme de réseau récurrent, capable de produire un déficit dans le traitement prosodique temporel des événement rapides, engendrant à la fois des difficultés pour une catégorisation lexicale et pour des tâches auditives temporelles de discrimination basées sur des tons purs (cf. Figure 6.1).

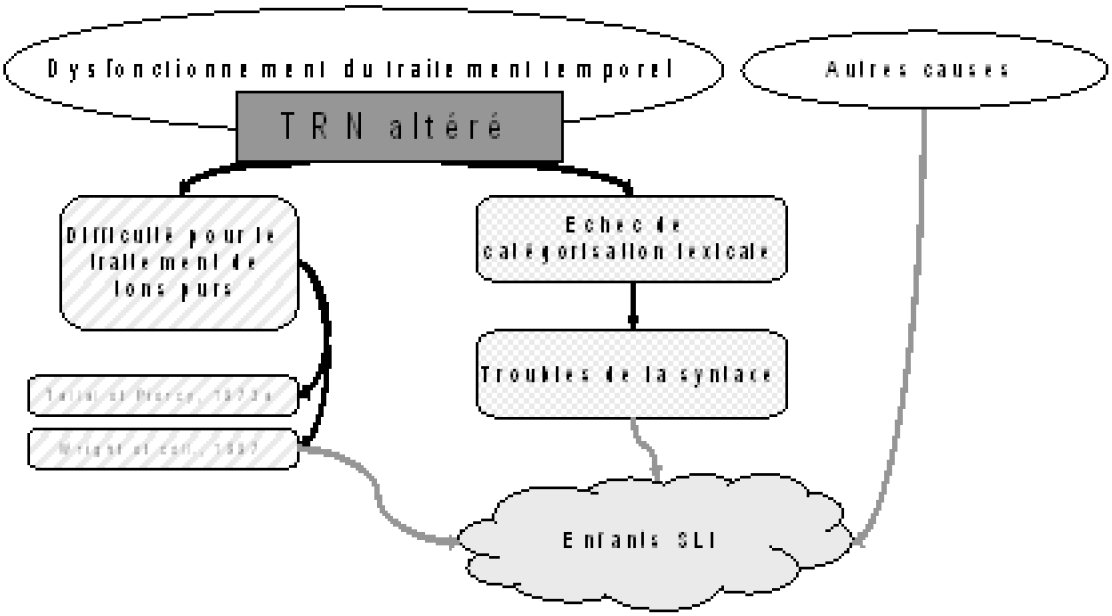


Figure 6.1 Diagramme de l'implication d'un déficit de traitement temporel auditif.

Pour ce faire, quatre expériences seront étudiées sur une population de 50 réseaux contrôles et altérés :

1. La première devra montrer les difficultés des réseaux altérés à discriminer les mots de fonction des mots de contenu. Elle permettra d'effectuer le réglage des constantes de temps pour obtenir les modèles altérés.
2. La seconde expérience prouvera que les réseaux altérés ont toujours la possibilité de catégoriser des séquences non linguistiques dont les éléments sont de longues durées.
3. La troisième expérience sera une réplication avec le réseau TRN de l'expérience de Tallal et Piercy (1973a) conduite avec des enfants SLI. Cette expérience montre que les enfants SLI sont perturbés pour le traitement des tons purs quand ils se suivent rapidement.
4. La dernière expérience modélisera l'expérience de Wright et coll. (1997). Celle-ci montre que les enfants SLI ont des difficultés uniquement dans le cas du masquage rétrograde (Backward Masking).

V.1.Catégorisation lexicale perturbée

La première expérience réalisée avec le TRN doit rendre compte des difficultés des enfants SLI pour la syntaxe. Nous postulons que les difficultés pour appréhender la syntaxe proviennent d'un échec de la reconnaissance des mots de fonction et de contenu, catégories grammaticales de bases. Nous choisissons donc d'affaiblir les performances de la tâche d'identification lexicale présentée dans le chapitre précédent. Cette section ne

tient compte que du corpus LSCP Français et de la segmentation en groupes de mots de fonction ou de contenu.

Maintenant, nous allons décrire les expériences tentées pour réduire cette classification. Pour pouvoir diminuer la résolution temporelle du réseau, nous choisissons d'augmenter les constantes de temps des unités du réseau.

Dans un premier temps, nous perturbons uniquement la couche de contexte ($State_D$), jusqu'à supprimer les connexions entre ces deux couches. Dans ces conditions, les performances sont très peu altérées. Ceci était prévisible, puisqu'une petite portion de la fin des groupes de mots (<100 ms) suffit pour les identifier (*cf.* Chap. 6 section IV.2.1*)*. Ainsi, cette information peut être traitée uniquement par les neurones de la première couche cachée (State). Nous avons vu que les réseaux TRN pouvaient utiliser la représentation MOMEL pour distinguer les catégories lexicales. Or, cette représentation ne contient pas d'indice micro-prosodiques. Nous avons donc testé l'influence des constantes de temps de la couche de contexte sur la population de réseau. Là encore, les performances sont peu influencées par cette modification (légère diminution de 68 % à 67 %). Dans ce contexte, la catégorisation effectuée par le réseau requiert un traitement local qui est assuré par la couche State du réseau. La couche StateD est loin d'être inutile, car elle est indispensable lors du traitement des structures globales. En outre, elle ne pertube pas la catégorisation des structures locales.

Nous considérons alors la première couche d'activation du réseau (State). Les constantes de temps sont augmentées (de 0.01s pour la population normale à 0.5s pour la population altérée). Nos premiers résultats ont permis de trouver une performance proche du hasard. Une première population de 50 réseaux (contrôle) distingue les catégories lexicales (F/C) avec une moyenne de 75 % correct sur le corpus de LSCP. Pour diminuer la sensibilité à la structure temporelle dans le TRN nous avons augmenté les constantes de temps du réseau pour produire 50 réseaux altérés. Les réseaux perturbés ne pouvaient accomplir correctement la même tâche (moyenne de 50 %, et maximum pour 53 % [114] ; Blanc et Dominey, 2001 et Blanc et coll., 2003 a&b).

La modification des constantes de temps de la couche State entraîne donc la perte des facultés nécessaires à l'identification des catégories grammaticales (Figure 6.2). Ces résultats illustrent donc les difficultés que peuvent éprouver les enfants SLI avec le traitement des mots de fonction. Dorénavant, nous n'utiliserons que l'augmentation des constantes de temps de la couche State. Ainsi, les constantes de temps des réseaux altérés ont une valeur de 0.5, alors que les réseaux contrôles ont une valeur de 0.01.

[114] Les performances augmentent de 50% à 60% si les réseaux sont initialisés au début de chaque phrase. Dans ce cas, le premier mot de fonction de chaque phrase est reconnu car il produit un vecteur nul dans le réseau. C'est pourquoi la moyenne des performances des réseaux les altérés se situe autour de 60 %.

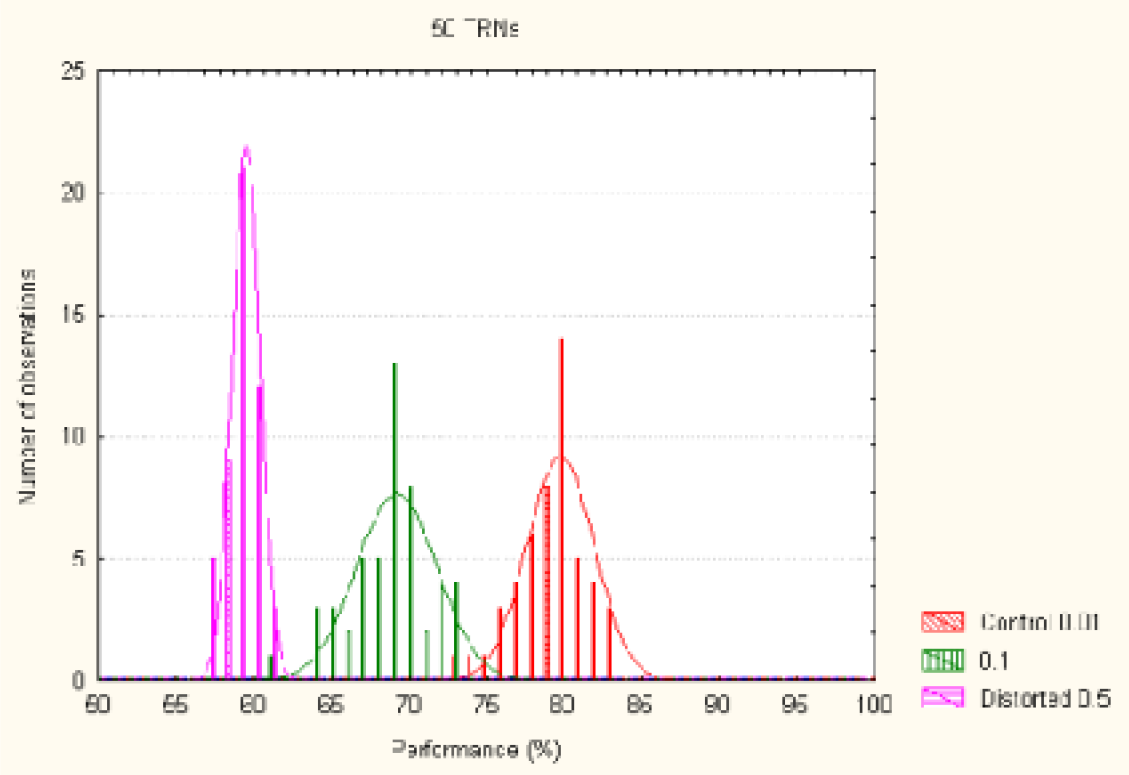


Figure 6.2 Répartition des performances pour la tâche de discrimination lexicale pour trois populations différentes de réseaux. Les réseaux sont initialisés au début de chaque phrase. (constante de temps de State= 0.01, 0.1, 0.5 ; Blanc et Dominey, 2002).

V.2.Identification de séquences constituées d'éléments de longue durée

Le trouble du traitement auditif des enfants SLI ne doit pas être un handicap global, dans la mesure où ils parviennent à maîtriser d'autres tâches cognitives. Ainsi, il faut vérifier si les réseaux, une fois perturbés, peuvent encore executer une tâche d'identification de séquences avec des éléments de longue durée. Cette expérience porte sur 3 séquences de 3 éléments stables notés A, B et C d'une durée de 800 ms présentés dans l'ordre suivant : ABC, CBC et CBA. Chaque élément a la même durée et active un seul neurone de la couche d'entrée. Les séquences peuvent êtres confondues par les deux premiers ou les deux derniers éléments. Dans le cas, où le réseau est sensible au dernier ou au premier élément, le taux d'identification est de 67 %, puisqu'il répond au hasard pour deux des séquences confondues. Les taux d'identification sont de 100 % pour la population contrôle, et de 99,33 % pour la population altérée, ce qui signifie que seul un réseau parmi 50 a confondu deux séquences entre elles, alors que les constantes de temps de la couche State ont été modifiées. Ainsi, les réseaux altérés n'ont pas de difficultés particulières pour identifier des séquences, dont les éléments ont de longues durées.

Tableau 6.1 Performance de la population de 50 réseaux TRN sans connexions récurrentes pour l'identification de séquences abstraites en fonction de la durée des éléments composant une séquence.

Durée d'un élément	% Correct
40 ms	100 %
80 ms	1 seul réseau à 100 %
100 ms	66.67 %
200 ms	66.67 %

Nous proposons de vérifier l'importance des connexions récurrentes pour identifier ces séquences, pour la population contrôle. Sans les connexions récurrentes, une séquence est systématiquement confondue avec une autre (Tableau 6.1). Mais les motifs d'activation sont quand même très proches pour les deux séquences identifiées correctement (distance << 1, chacune des 25 unités du réseau traite des valeurs entre 0 et 100). Les connexions récurrentes sont donc indispensables pour identifier des séquences dont les événements discriminants sont éloignés dans le temps, i.e. forment une structure globale. Cependant si ceux-ci sont proches, la mémoire provenant des intégrateurs à fuite eux-même est suffisante. Même si les prototypes pour ces séquences simples sont proches, il semble que cette capacité du réseau soit suffisante pour identifier certain mots de fonction / contenu. Cette particularité permet de comprendre pourquoi la parole a recours a tant d'évènements de courtes durées. Effectivement l'apprentissage des indices rapides est simplifié dans la mesure où il n'est fait appel ni à des connexions supplémentaires ni à une couche supplémentaire. Les événements de courtes durées sont donc moins coûteux à traiter.

Les réseaux TRN altérés exhibent une certaine incapacité pour distinguer les mots de fonction des mots de contenu, bases des structures grammaticales, tout en répondant correctement à une tâche simple d'identification de séquences avec des élements de longue durée. Est-il possible d'obtenir les mêmes performances que les enfants SLI avec des tâches de discrimination de stimuli auditifs rapides ?

V.3.Tâche de perception auditive rapide

La tâche originale sera d'abord décrite avant d'être modélisée par le réseau TRN.

V.3.1.Tâche originale

La tâche est décrite dans l'article de Tallal et Piercy paru en 1973a. Elle a été réalisée sur une population d'enfants SLI, et d'enfants contrôles qui ne présentent pas de troubles du langage. Cette tâche a été également reprise dans un article récent avec des nourrissons de 7 mois et demi (Benasich et Tallal, 1996), et une modélisation avec des rats (Clark et coll., 2000).

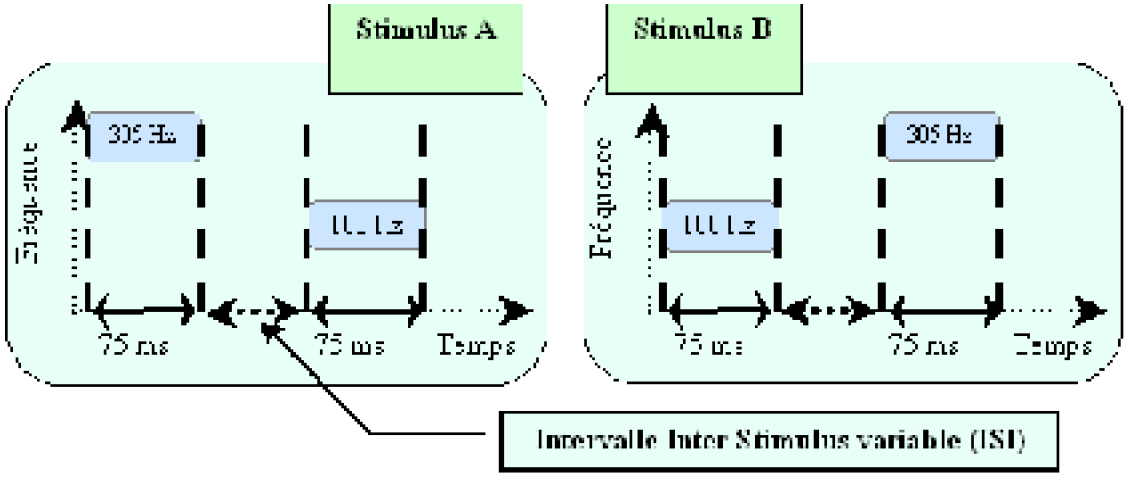


Figure 6.3 Illustration des deux types de stimuli utilisés dans la tâche de Tallal et Piercy (1973a).

Deux stimuli auditifs doivent être discriminés. Ils diffèrent uniquement par l'ordre de deux tons purs d'une durée de 75 ms avec deux fréquences fondamentales distinctes (100 Hz et 305 Hz). L'intervalle entre les deux tons purs (noté ISI) varie entre 8 et 3543 millisecondes.

Les enfants SLI discriminent avec difficulté les stimuli pour un seuil ISI inférieur à 300 ms, alors que les enfants normaux ne rencontrent pas de problèmes particuliers (Figure 6.5).

V.3.2.Simulation de la tâche

Pour pouvoir effectuer cette tâche, il faut seulement que les réseaux fournissent des vecteurs distincts pour chacun des stimuli. Dans le cas contraire, le réseau ne pourra pas distinguer les stimuli.

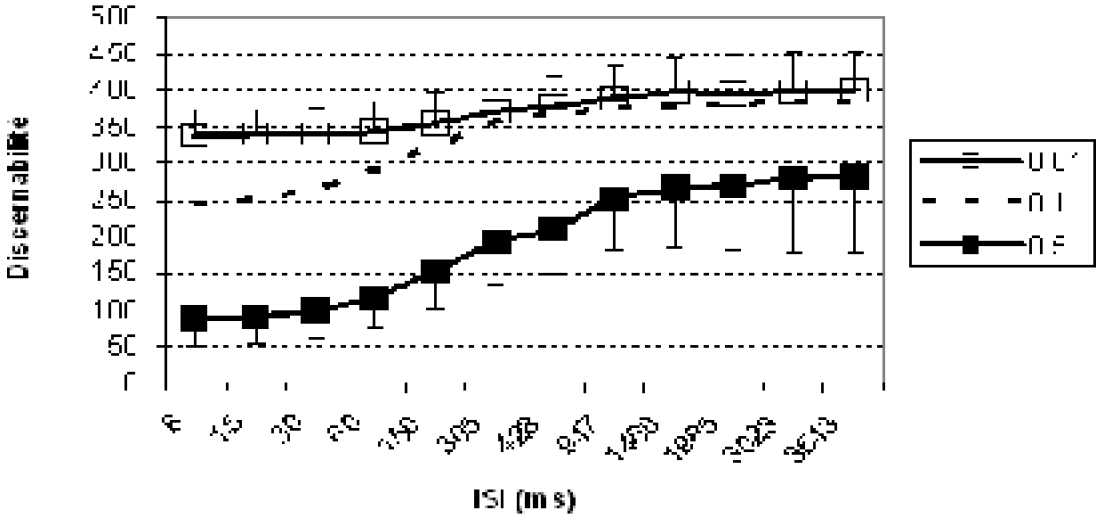


Figure 6.4 Discernabilité des stimuli pour la tâche de discrimination des tons purs

(moyenne obtenue sur 50 réseaux, les écart types sont indiqués par les barres verticales). La durée entre les deux stimuli est présentée en abscisse en millisecondes (ISI). Trois populations avec des constantes différentes (contrôles 0.01, 0.1 et altérés 0.5) sont représentées.

Notre approche se fera en trois temps. En premier lieu, nous indiquerons les distances obtenues entre les vecteurs de chaque stimuli, en fonction de la durée ISI et aussi des valeurs des constantes de temps des réseaux (contrôles : 0.01 et altérés : 0.5). Ensuite, nous testerons l'impact d'une modification de la constante de temps, pour un seuil d'apprentissage donné. Enfin, nous indiquerons les valeurs obtenues par les réseaux contrôles, lorsque le seuil d'apprentissage est trop élevé, i.e. un entraînement insuffisant est simulé.

La section précédente nous a enseigné que l'identification lexicale était perturbée lorsque les constantes de temps étaient augmentées. En effectuant ce seul changement nous ne constatons pas de dégradations des résultats, pour l'identification des stimuli auditifs. En revanche, un indice de discernabilité des stimuli A et B peut être obtenu en calculant la distance euclidienne entre le vecteur du stimulus A et celui du stimulus B. Cette distance est indiquée dans la figure 6.4.

Ainsi il est possible de simuler les résultats observés par Tallal et Piercy (1973a) en fixant un seuil en dessous duquel deux vecteurs ne peuvent être distingués. Les motifs obtenus pour les stimuli sont plus proches pour les réseaux altérés que pour les contrôles, quelle que soit la durée de l'intervalle ISI. Le modèle donne également des représentations des stimuli A et B plus proches lorsque l'intervalle entre les deux tons est bref, et cette proximité est encore plus importante pour les réseaux altérés. Il s'ensuit que les réseaux altérés ont plus de difficultés pour discriminer les stimuli contenant un intervalle bref, que les réseaux contrôles.

Il ne reste donc qu'à choisir une valeur de seuil. Celui-ci devra permettre de refléter les résultats observés avec les enfants SLI et contrôles. De plus, ce seuil permet de simuler un apprentissage progressif des stimuli. Une façon un peu plus complexe de le simuler serait d'employer un réseau supplémentaire (mémoire associative, carte de Kohonen) pour apprendre les vecteurs donnés par le réseau TRN. Les cycles d'apprentissage successifs correspondraient à une diminution progressive de ce seuil. Il a été déjà été démontré que des enfants SLI peuvent répondre correctement à cette tâche, après un entraînement plus long que celui des enfants normaux (Tomblin et Quinn, 1983 ; Tallal et coll., 1981).

Les réseaux altérés ont en moyenne plus de mal à faire la distinction entre les stimuli pour des intervalles inférieurs à 150 ms (Blanc et Dominey, 2002 ; Blanc et coll., 2003 a&b). Pour retrouver la valeur de 300 ms annoncée par Tallal et Piercy (1973a), il faudrait rechercher les constantes de temps et le seuil d'apprentissage du réseau de façon systématique.

Nous souhaitons maintenant examiner le comportement des réseaux contrôles lorsque le seuil d'apprentissage est fixé à une valeur trop élevée, simulant ainsi un entraînement insuffisant. Ainsi, le déficit est simuler par un manque d'apprentissage et non par un défaut du mécanisme de perception. Dans ce cas, les performances croissent

progressivement, et le schéma de réponse des enfants SLI n'est pas retrouvé (Figure 6.7). Effectivement, les réponses des réseaux augmentent progressivement et non brutalement entre 150 ms et 425 ms.

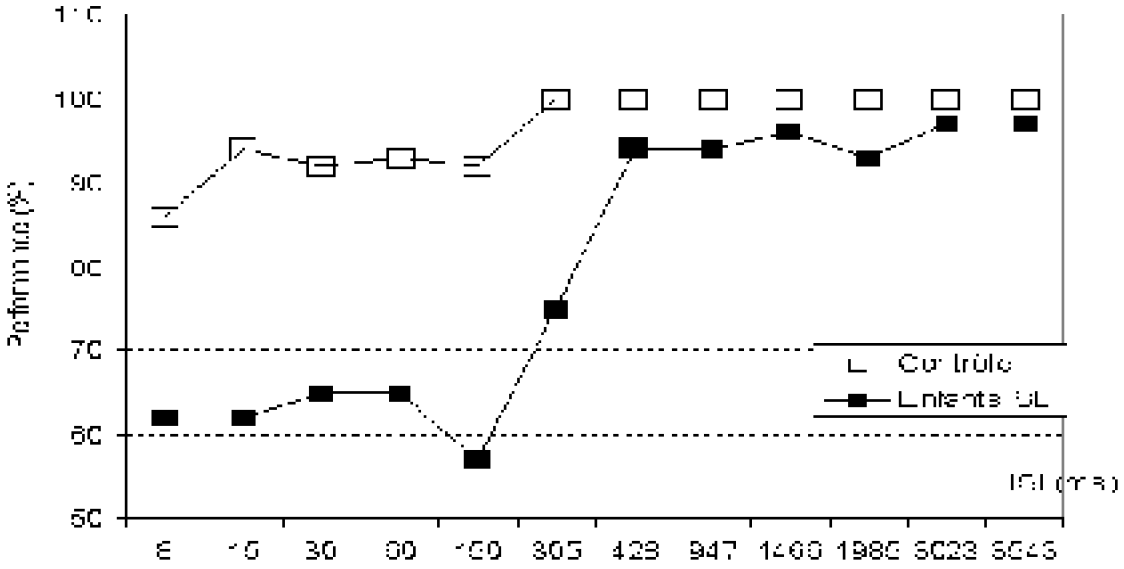


Figure 6.5 Performances des enfants (SLI et contrôle) pour la tâche de perception rapide (Tallal et Piercy, 1973a).

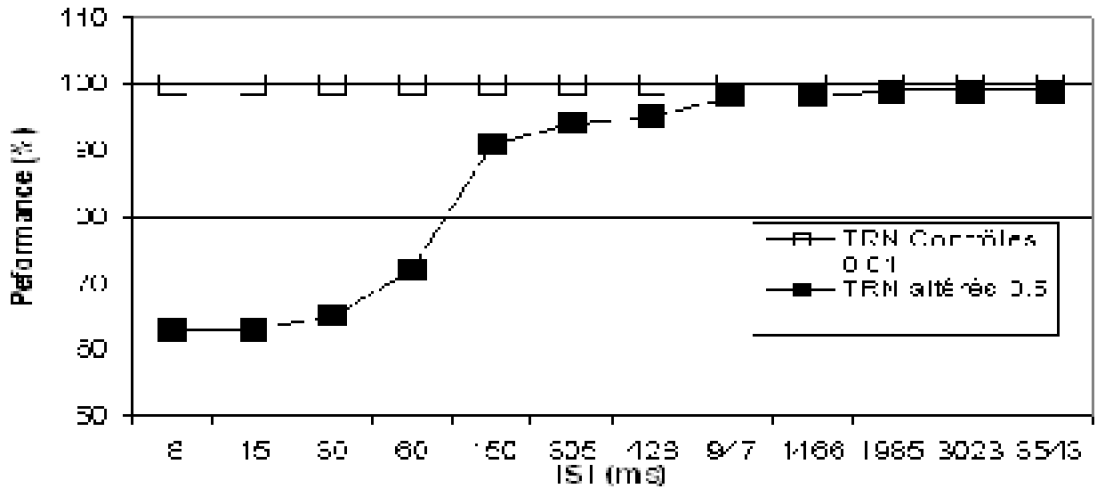


Figure 6.6 Performances des deux populations de 50 réseaux TRN (altérés et contrôles) pour la tâche de perception rapide (Tallal et Piercy, 1973a) avec un seuil fixé à 150.

Ainsi, pour simuler le comportement des enfants SLI, il faut tenir compte d'un seuil d'apprentissage (simulant l'entraînement) et d'une constante de temps augmentée (pour modéliser le trouble de traitement).

Des études expérimentales complémentaires ont précisé que les déficits de traitement rapide des enfants SLI apparaissent lors du masquage d'un ton pur par un bruit. La modélisation par le réseau TRN peut-il permettre de retrouver ce phénomène ?

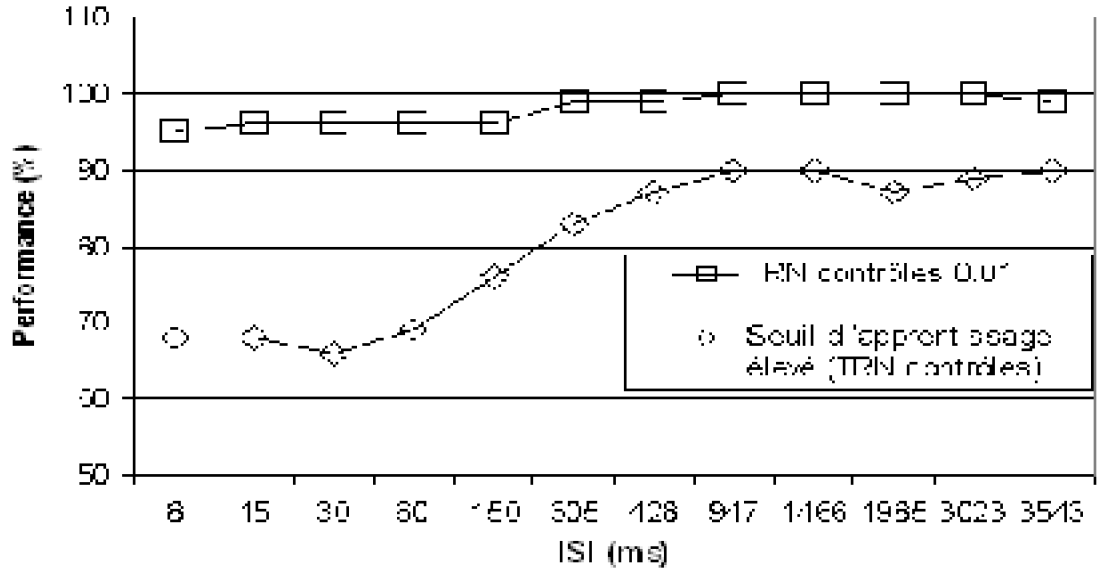


Figure 6.7 Performance d'une population contrôle (constante de temps 0.01) de réseaux TRN soumis à deux seuils différents pour l'apprentissage.

V.4.Tâche de masquage auditif

Comme pour la section précédente, l'expérience originale (Wright et coll., 1997) sera détaillée, avant de proposer deux types de modélisations de cette tâche. La première modélisation emploiera une représentation abstraite des données sensorielles transmises au réseau TRN. La seconde se fondera sur la description obtenue par le cochléogramme, à partir des sons utilisées dans l'expérience originale.

V.3.1.Tâche originale

Nous proposons maintenant d'étudier une autre tâche effectuée avec des enfants contrôles et SLI (Wright et coll., 1997). Elle a été initialement conçue à partir d'une hypothèse qui stipulait que les voyelles pouvaient agir comme des masques sur les consonnes (Tallal, Stark, Kallman et Mellits, 1981). Wright et son équipe (1997) ont montré que les enfants SLI avaient besoin de plus d'intensité pour percevoir un ton pur suivi d'un masque de bruit. Cette expérience pourrait suggérer une atteinte du mécanisme sensible aux fréquences, dans la cochlée à la périphérie du système auditif. Cependant, ce déficit n'apparaît pas lorsque le ton pur est présenté après le bruit.

Cette tâche emploie deux stimuli : un ton pur (1000 Hz) qui dure 200 ms dans la condition contrôle, et 20 ms dans toutes les autres conditions , ainsi qu'un bruit blanc (masque) qui dure 300 ms. Ce dernier apparaît sous deux formes : une condition où toutes les fréquences entre 600 Hz et 1400 Hz sont activées (bruit passe bande), et une autre ou les fréquences proches de celle du ton pur ne sont pas présentées (notched noise : fréquences activées de 400 à 800 Hz et de 1200 à 1600 Hz).

Cette tâche est composée de cinq conditions différentes. La condition 1.contrôle utilise un ton pur de longue durée (200 ms), que les enfants doivent percevoir, alors que

le masque est émis en même temps que le ton pur. Les quatre autres conditions font appel à un ton pur bref de 20 ms, qui est placé selon quatre manières différentes par rapport au masque : juste avant (2. Backward), en même temps dès le début du masque (3. Simul-onset), en même temps au milieu du masque (4. Simul-delay), ou après le masque (5. Forward Masking, cf. Figure 6.8).

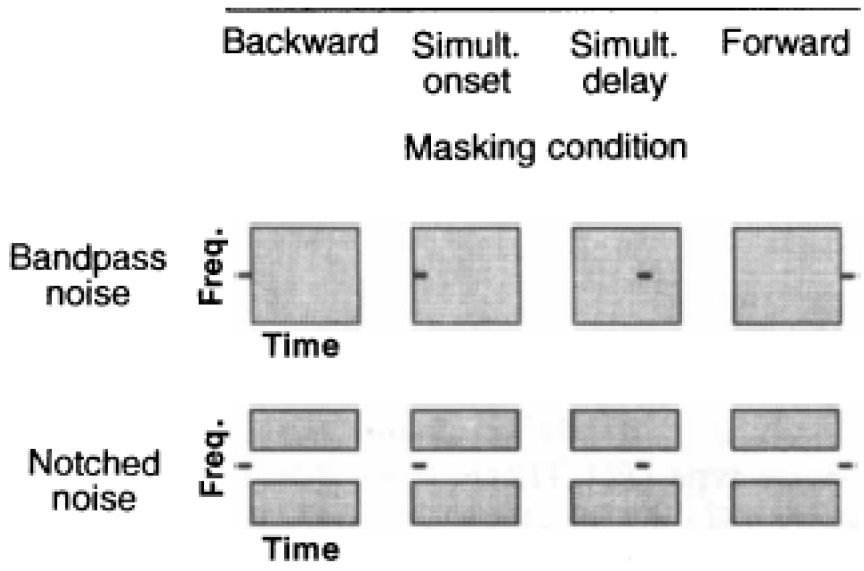


Figure 6.8 Illustration des différentes conditions de la tâche de masquage, tirée de Wright et coll. (1997)

La condition contrôle ne présente clairement aucune différence entre les enfants. En revanche, les enfants SLI ont besoin d'une intensité du ton pur plus importante que les enfants contrôles pour percevoir le ton pur. Toutefois, cette différence n'est significative que pour la condition de masquage rétrograde (Backward Masking, Figure 6.9).

Nous allons maintenant proposer deux types de représentation pour transmettre les données au réseau TRN. La première est abstraite et suit celle de F0 utilisée pour les attitudes prosodiques (*cf.* Chap. 4 section III.3). La seconde est utilisée avec le cochléogramme qui analyse des stimuli sonores (testé pour l'IAL, et l'identification des mots de fonction et de contenu).

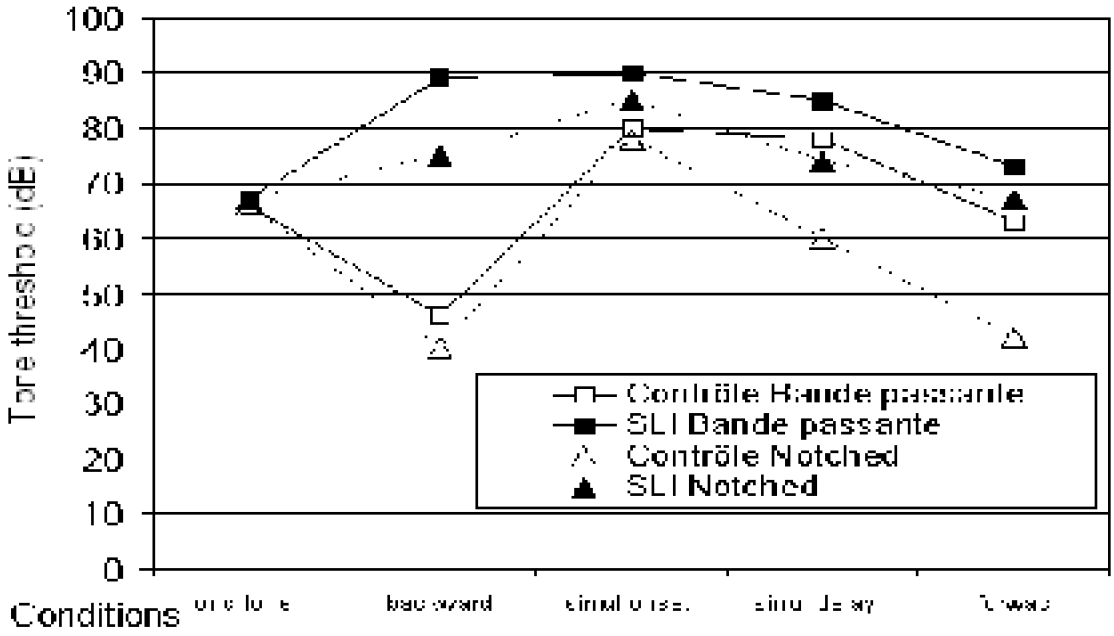


Figure 6.9 Seuil de discrimination du ton pur dans les deux conditions de bruit (Bande passante et Notched) et les deux groupes d'enfants SLI et contrôle (d'après Wright et coll., 1997).

V.3.2.À partir d'une représentation abstraite

Concrètement le ton pur sera représenté par l'activation variable d'un seul des neurones consacrés à la fréquence fondamentale, tandis que le bruit sera traduit par l'activation à 40 % de l'ensemble des autres neurones d'entrées. Pour chaque expérience, le ton pur voit son intensité varier entre 0 et 100, de 10 en 10. Si le ton pur n'est jamais identifié par un réseau, ses performances ne seront pas prises en compte. Avec cette représentation, il est impossible d'identifier le ton pur, lorsqu'il est mélangé dans le bruit, tant que son intensité ne dépasse pas celle du bruit.

Deux cas seront examinés :

1. Seuls les réseaux capables de répondre aux deux premières conditions sont considérés. Dans ce cas, nous montrons uniquement que le réseau TRN est sensible aux conditions de traitement rapide, lié à la durée du ton pur.
2. Seuls les réseaux capables de fournir une réponse aux cinq conditions de la tâche seront conservés ;

V.3.2.1.Deux premières conditions seules

Les seuils d'intensité sont présentés dans la figure 6.10. Comme chez les enfants la tâche contenant le ton long est effectuée avec la même intensité pour le ton pur chez les deux groupes, alors que la condition « Backward Masking » requiert une intensité plus élevée pour les réseaux altérés. La troisième colonne de la figure 6.11 indique le pourcentage

d'erreurs effectuées par les réseaux contrôle (State : 0.01) et altérés (State : 0.5) accomplissant les deux tâches de perception auditive, et l'identification des mots de fonction et de contenu. Il résulte que les réseaux récurrents ne pouvant effectuer la tâche d'identification grammaticale ont plus de difficultés pour percevoir le ton pur dans la tâche de « Backward Masking » (Blanc et Dominey 2001 et Blanc et coll., 2003).

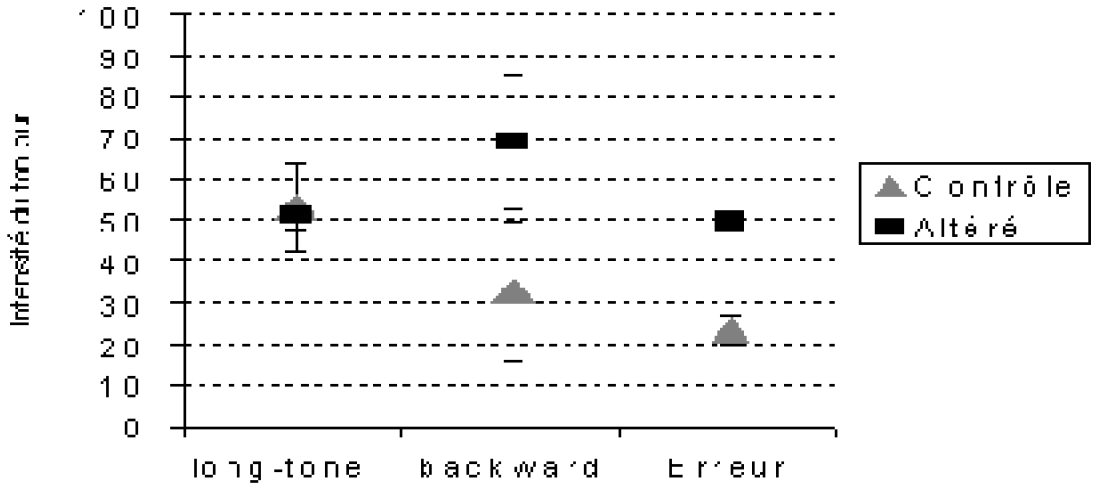


Figure 6.10 Discrimination de séquences (condition Long-tone, Backward Masking et pourcentage d'erreur de la tâche d'identification des mots de fonction et de contenu). Seules les deux conditions Long-tone et Backward Masking ont été étudiées.

V.3.2.2.Les cinq conditions

Lors de la simulation de la tâche avec les réseaux, il apparaît que certains d'entre eux ne sont pas capables de réaliser la tâche pour toutes les conditions. Dans ce cas, il n'est pas possible de déterminer un seuil d'intensité suffisant pour percevoir le ton pur pour certains réseaux. Ainsi, parmi la population des 50 réseaux altérés, seulement 2 réseaux sont capables de faire les cinq conditions de la tâche, contre 15 réseaux contrôles.

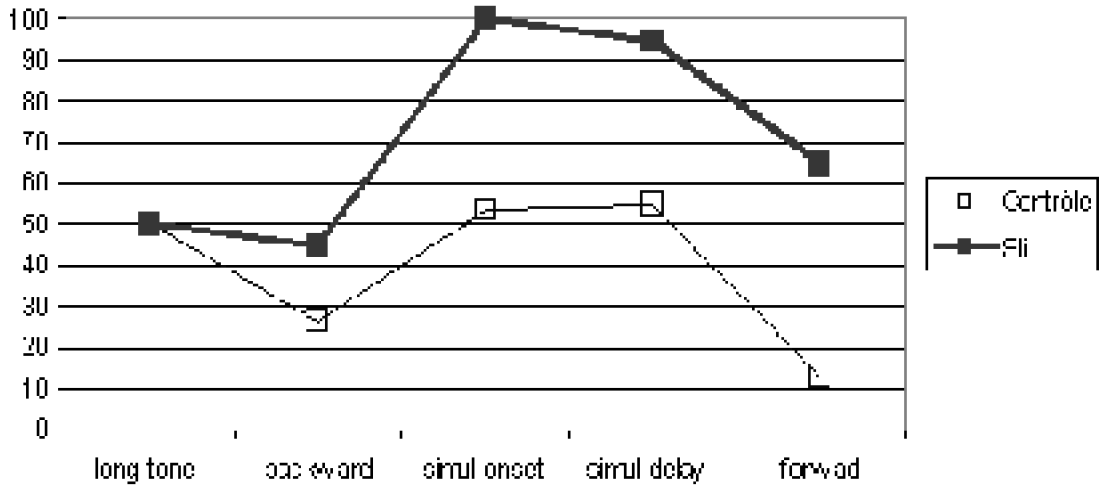


Figure 6.11 Seuil d'identification avec une représentation abstraite des stimuli auditifs, pour une population contrôle et altérée (SLI) des réseaux TRN.

Le graphique 6.11 illustre les seuils observés pour les réseaux. Nous retrouvons l'allure générale des courbes obtenues pour les enfants. Cependant, la différence entre les deux courbes altérées et contrôles est plus importantes qu'entre les enfants normaux et SLI. Seule la condition Long-tone donne le même seuil pour les deux groupes.

Pourquoi y a-t-il si peu de réseaux en mesure de faire la tâche alors qu'ils arrivent tous à faire la tâche d'identification F/C ?

Il est probable que les choix faits pour la représentation des stimuli ne soient pas les bons. Le chapitre précédent (Chap. 4 section IV.1) a révélé que le paramètre sigma qui détermine le nombre de neurones activés, ainsi que le nombre de neurones codant F0 influence les performances. En particulier, nous avons vu que certaines configurations permettaient de réduire l'espace des performances d'une population de réseaux.

Effectivement, plus il y a de neurones de la couche d'entrée activés, plus nombreux sont les réseaux en mesure d'effectuer l'ensemble des 5 conditions de la tâche. En outre, cette même modification entraîne que l'intensité moyenne nécessaire baisse principalement pour les conditions simultanées (3 et 4). Cependant, un autre paramètre du codage nous semble discutable : l'intensité. Effectivement, la sensation de l'intensité se mesure sur une échelle linéaire, alors que l'énergie, son corrélat physique suit une échelle exponentielle [115] . Le premier codage étudié s'appuie donc des impressions perceptives plus que sur les données physiques. Cependant, si l'échelle d'intensité est traduite de manière exponentielle, il semble que le nombre de réseaux percevant le ton pur devrait augmenter. Le codage du son par le cochléogramme reprend les propriétés qui viennent d'être abordées.

[115] Le niveau sonore correspond à la sensation de volume sonore. Il se mesure en bels et décibels (dB). Le bel est une échelle logarithmique d'intensité. Le décibel (dB) est le 1/10 d'un bel. On calcule les dB à partir de l'intensité, suivant la formule dB SPL = 10 log10 (I/Ir) = 20 log10 (P/Pr).

V.3.3.À partir du cochléogramme

Nous choisissons de reprendre le codage du son par cochléogramme [116] étudié dans les chapitres IAL et de catégorisation lexicale. Cette représentation est fondée sur des données issues de la psychoacoustique.

La première difficulté est de fixer l'intensité du bruit. Nous faisons en sorte de rester au même niveau que précédemment. Ainsi, l'activation maximale des unités du réseau est 40 pour le bruit. Mais chaque unité possède une activation différente. En outre, le ton pur peut être détecté même lorsque son intensité est très faible, quand les deux stimuli sont comparés. Comme pour l'expérience précédente (section IV.3), un seuil sera appliqué pour simuler un apprentissage. Ainsi, les vecteurs générés pour les stimuli incluant le ton pur seront différents de celui ne contenant pas le ton pur, mais ils pourront être confondus en dessous d'un seuil d'apprentissage fixé. Le seuil est fixé à 10. Cette valeur a été obtenue après sélection, de façon à ce que le profil des réponses reproduit le profil observé chez les enfants SLI, avec un seuil d'apprentissage commun aux réseaux altérés et contrôles.

L'utilisation de cette nouvelle représentation permet qu'un plus grand nombre de réseaux puisse répondre aux cinq conditions de la tâche. Sans le seuil d'apprentissage, les seuils d'intensité restent identiques entre les réseaux contrôles et les réseaux altérés. Cependant, l'application du seuil fait que certains réseaux ne répondent plus aux cinq conditions. Dans ce cas, la valeur d'intensité nécessaire est considérée comme étant supérieure à 110 dB.

Lorsque les 50 réseaux sont pris en compte, les réseaux altérés et contrôles ne diffèrent jamais sur la tâche contrôle avec le ton long, quel que soit le seuil d'intensité. Mais les deux populations se distinguent sur toutes les conditions, où le ton pur dure seulement 20 ms. Ainsi, les réseaux sont donc bien sensibles aux durées des stimuli. Lorsqu'ils sont altérés, ils ne perçoivent plus le ton pur lorsque celui est bref. Cependant, si les quatre conditions où est présent le ton bref (20 ms) sont comparées, la plus grande différence entre les réseaux contrôles et altérés apparaît pour la tâche Forward, et non pas pour la pour la tâche Backward Masking. Comparés aux êtres humains, les réseaux ont un seuil d'intensité assez élevé pour discerner le ton pur bref dans la tâche Backward (5 dB de différence entre la condition Long-tone et la condition Backward pour les réseaux contrôles contre 20 dB pour les enfants contrôles, Figure 6.12).

[116] Nous n'avons pas utilisé PRAAT, mais la toolbox développée par M. Slaney sous Matlab, et la représentation de Meddis.

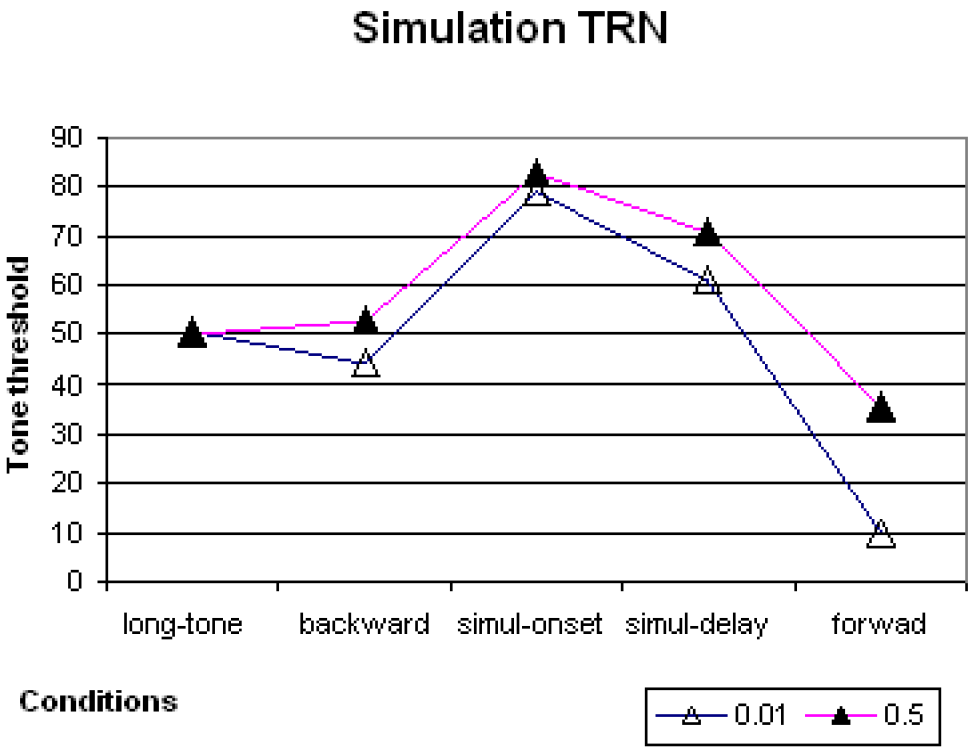


Figure 6.12 Seuil d'intensité pour les cinq conditions de la tâche. Moyenne des 50 réseaux TRN contrôles (0.01) et altérés (0.5)

Nous suggérons alors de ne garder que les réseaux contrôles effectuant la tâche Backward avec un seuil inférieur à la condition contrôle, et de les comparer aux mêmes réseaux altérés (Figure 6.13). Malgré ces dispositions l'écart entre les réseaux contrôles et altérés reste important pour la condition Forward. Il est possible que la représentation du cochléogramme ne tienne pas assez compte du phénomène de double masquage. Ce phénomène de double masque ne peut se produire uniquement que dans la condition Bandpass. Ainsi, bien que les données fournies au réseau soient issues de la condition Bandpass, le profil des performances ressemble plus à celui de la condition Notched, qui supprime cet effet de double masquage.

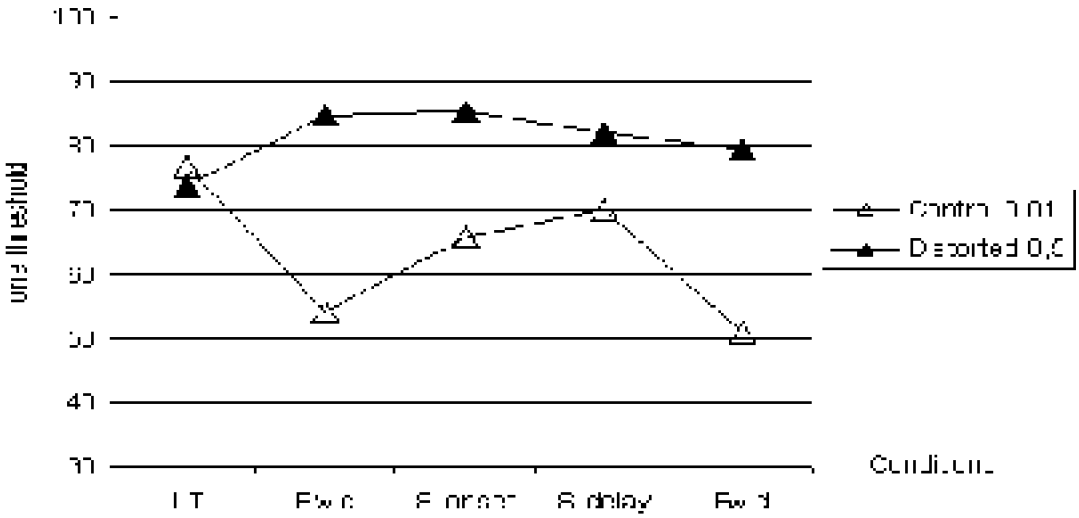


Figure 6.13 Simulation de la tâche de Wright et coll. (1997) (Blanc et Dominey, 2002). Seuls les réseaux contrôles effectuant la tâche Backward avec un seuil inférieur à la condition Long-tone sont conservés, ainsi que les réseaux altérés correspondant à ces réseaux. Les cinq conditions sont Long-tone, Backward, Simultaneous-onset, Simultaneous-delay et Forward.

Comme pour la tâche précédente de discrimination auditive rapide, nous suggérons d'étudier l'influence d'un seuil de discrimination élevé, pour simuler le facteur d'apprentissage. Dans ce contexte, les constantes de temps du réseau restent inchangées (Figure 6.14).

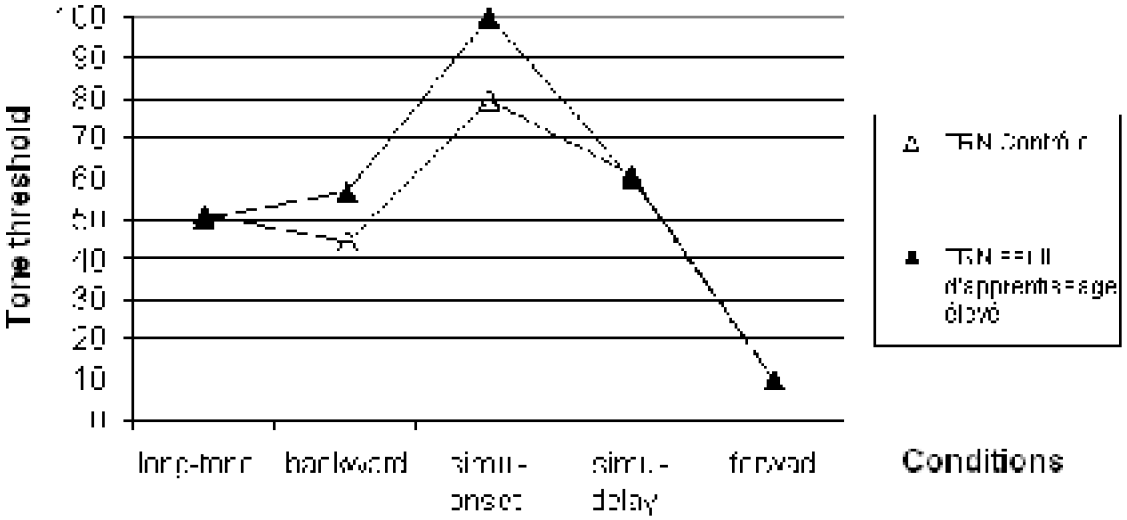


Figure 6.14 Simulation de la tâche de Wright et coll. (1997) pour 50 réseaux. Les enfants SLI sont simulés avec un seuil d'apprentissage trop élevé (fixé à 20) par rapport au réseau contrôle (fixé à 5), ce qui s'interprète comme un manque d'entraînement.

Cette fois-ci, les conditions Forward et Simul-delay ne permettent pas de distinguer les réseaux contrôles des réseaux altérés, alors que les conditions Backward et Simul-onset les distinguent. En conclusion, il semble que pour pouvoir répliquer les résultats de Wright et coll. (1997) les constantes de temps, (i.e. diminuer les capacités de

traitement temporel rapide) et le seuil de discrimination (i.e. diminuer l'entraînement pour la tâche) doivent être augmentées.

Finalement notre modèle semble rejoindre les principales critiques effectuées pour la tâche de Wright (1997). Seul un manque d'apprentissage semble pouvoir expliquer la différence entre les modèles contrôles et altérés pour la condition Backward Masking, alors que des constantes de temps élevées différencient les populations dans toutes les conditions où intervient le ton pur bref.

VI.Discussion

Pour expliquer les troubles linguistiques des enfants SLI, des études comportementales ont démontré des déficits perceptifs pour les événements de courtes durées. Si certaines critiques ont été émises sur la systématicité du lien entre les troubles manifestés par les enfants SLI, et ces tests auditifs, il s'avère que ceux-ci sont liés principalement aux capacités de compréhension grammaticale (Bishop et coll., 1999) Il semble ainsi indéniable que les difficultés pour comprendre ces structures sont en relation avec une incapacité à traiter des événements brefs ou présentés rapidement, au moins pour un sous-groupe d'enfant SLI.

En concevant un système respectant une contrainte temporelle, est-il possible d'exhiber les mêmes difficultés pour des tâches s'appuyant à la fois sur du matériel verbal et non-verbal ?

Pour évaluer cette hypothèse, nous avons montré dans un premier temps qu'un réseau récurrent sensible à la structure temporelle locale est capable d'effectuer cette discrimination à partir du contour prosodique. Puis, nous avons prouvé que, suite à une modification des constantes de temps du réseau, celui-ci devient moins approprié pour exécuter cette catégorisation, alors même qu'il reste performant pour des séquences de plus longues durées. La structure prosodique locale n'est plus décodable par le TRN à partir des séquences temporelles. Privé de cette discrimination fondamentale, le système ne peut parvenir à développer facilement une syntaxe intacte. Ces démonstrations soutiennent 1) que la prosodie fournit bien des informations nécessaires pour l'identification des deux catégories grammaticales, et 2) qu'un système privé de sensibilité pour les structures temporelles ne peut exploiter ces informations, comme c'est le cas chez les enfants SLI.

Dominey et coll. (2003) proposent une architecture qui nécessite une telle catégorisation pour comprendre des structures syntaxiques. Dans un premier temps, nous discuterons des résultats obtenus, avant de traiter de l'implication d'un modèle informatique pour montrer le lien entre la syntaxe et les évènements auditifs rapides.

VI.1.Les expériences

Le premier point concerne le lien entre la syntaxe et le traitement temporel de séquences.

En particulier, nous avons examiné comment ce traitement peut être altéré, de façon à empêcher la discrimination de catégories lexicales tout en laissant intacte l'identification de séquences d'éléments de longues durées. Ensuite, nous présenterons les observations faites en neurophysiologie montrant des troubles relatifs à ce type de traitement. Nous discuterons alors de la modélisation des taches de Tallal et Piercy (1973a) et de Wright et coll. (1997). Enfin nous ferons le lien entre nos résultats et l'hypothèse du déficit de traitement rapide.

VI.1.1.Syntaxe et traitement temporel de séquence

La première expérience a permis de déterminer les paramètres à modifier pour empêcher toute catégorisation lexicale des mots à partir de l'intonation. Nous avons montré qu'il était possible de diminuer progressivement les performances de la population de réseaux, en augmentant les constantes de temps de la première couche cachée. Nous montrons ainsi que la valeur de ces constantes a une influence sur les résultats, contrairement à ce qui est annoncé dans les travaux de Buonomano et coll. (1995), avec d'autres modèle de réseaux temporels. Cependant, les valeurs de ces constantes de temps doivent être fortement modifiées pour s'assurer que tous les réseaux soient incapables de répondre à la tâche.

En outre, nous remontrons que le réseau travaille à partir d'une échelle locale pour effectuer la catégorisation lexicale, puisque les connexions récurrentes ne sont pas nécessaires. Cela avait déjà été examiné dans le chapitre précédent (section IV.2.1 Analyse du réseau). Nous prouvons donc que le réseau peut avoir des difficultés pour catégoriser les mots à partir de F0, suite à un dysfonctionnement des unités du réseau. Si le réseau ne peut procéder à cette catégorisation, il aura donc des difficultés pour maîtriser les éléments syntaxiques et, plus probablement, ceux les plus courts comme les mots ou phonèmes de fonction. Dans un contexte d'acquisition du langage, les réseaux devraient montrer les mêmes difficultés que les enfants SLI avec la syntaxe. Contrairement à notre simulation, les enfants SLI ont à leur disposition d'autres informations pour comprendre les catégories syntaxiques. C'est pourquoi, leur difficulté s'exprime sous la forme d'un retard.

VI.1.2.Aspect neurophysiologique

L'augmentation des constantes de temps du réseau simule un trouble biologique, lié aux neurones. Les recherches courantes en neurophysiologie ont maintenant montré qu'un trouble du traitement phonologique, qui est crucial pour le développement du langage écrit et oral, proviendrait, au moins en partie, de difficultés pour la perception et la production des informations sensori-motrices en succession rapide (de l'ordre de quelques millisecondes). Les données neurophysiologiques indiquent qu'un stimulus n'est pas correctement représenté au niveau neural préattentif dans les conditions de transitions rapides, chez les enfants SLI. Dans ce cas, les données neurophysiologiques coïncident avec les données comportementales. La discrimination comportementale est retrouvée dans les mesures électrophysiologiques, concernant le chemin de l'audition. Elle ne dépend pas de l'attention ou d'une réponse volontaire. Leurs résultats prouvent

que les difficultés des enfants SLI interviennent avant la perception consciente (Bradlow et coll., 1999).

Un dommage dans l'hémisphère gauche peut détruire le mécanisme responsable du traitement des changements acoustiques brefs et rapides, qu'il soit verbal ou non. Cette perturbation pourrait être à l'origine de certaines aphasies. Le traitement des informations les plus rapides serait localisé dans l'hémisphère gauche. Ainsi, l'évolution aurait permis aux être humains d'utiliser ce système pour traiter la parole, et tirer ainsi parti de différences acoustiques très brèves (Tallal et coll., 1998). Quelques études avec pour modèles des rats (Clark et coll., 2000), ou s'appuyant sur des analyses post-mortem de dyslexiques montrent que des malformations morphologiques jouent un rôle dans le déficit du traitement des informations rapides (Benasich et Tallal, 2002).

Nous avons établi que les réseaux TRN altérés n'éprouvaient pas de difficultés lors de l'identification de séquences d'éléments de longues durées. Nous montrons ainsi que le mécanisme de traitement n'est pas entièrement inopérant. Ainsi, une aberration au niveau de l'architecture du réseau peut perturber l'identification portant sur des éléments de courtes durées, alors que les séquences d'élèments de longue durée sont correctement identifiées. En outre, nous dévoilons que les connexions récurrentes entre les deux couches cachées State et $State_D$ sont indispensables pour traiter les séquences les plus longues (section V.2). La boucle récurrente entre les couches State et $State_D$ permet le traitement des informations globales, alors que la couche State traite les évènements à un niveau local.

VI.1.3.Tâche de discrimination rapide (Tallal et Piercy, 1973a)

Après avoir montré les difficultés des réseaux TRN altérés avec les catégories syntaxiques, nous montrons qu'ils reproduisent le même profil de réponses que les enfants SLI pour la tâche de discrimination temporelle mise au point par Tallal et Piercy (1973a). Cette tâche nous permet d'appuyer que le déficit de traitement temporel a des conséquences sur le traitement de la prosodie. Effectivement, la durée et la fréquence fondamentale sont les seules variables de cette tâche. D'autre part, l'intervalle de temps affecté par le déficit se situe autour d'une syllabe, ce qui corrobore l'idée que les mots de fonction ne pourraient être traités convenablement, dans la mesure où ceux-ci contiennent d'une part rarement plus de deux syllabes et d'autre part des syllabes plus rapides que celles des mots de contenu (Chapitre 5 section IV.1.1 et Shi et coll., 1998).

Récemment, cette tâche a été reprise avec des nouveau-nés (Benasich et Tallal, 1996). Les auteurs ont trouvé que si les nourrissons éprouvaient des difficultés lors de cette expérimentation, ils présenteront également des troubles du langage par la suite. Les nourrissons qui ont des seuils élevés pour effectuer cette discrimination (supérieurs à 150 ms) présentent des facultés diminuées pour le langage (Benasich et Tallal, 2002).

La perturbation provoquée sur la première couche de traitement (State) du TRN engendre sans doute les mêmes difficultés que si la fenêtre d'analyse des récepteurs était agrandie. Avec une fenêtre d'analyse plus large, il devient difficile de situer les événements dans le temps. Mais cette perturbation engendrerait des problèmes touchant d'autres domaines de l'audition, comme le traitement du rythme. Dans notre cas, nous

pouvons imaginer que c'est un seul mécanisme dédié à l'intonation qui est touché. Si les entrées ne sont pas altérées, seul le traitement de la F0 par ce réseau précis est endommagé, et donc la syntaxe serait touchée. Il est possible que ce mécanisme perturbé soit utilisé pour le traitement des indices rapides, auquel cas, il pourrait engendrer des problèmes touchant d'autres modalités.

VI.1.4.Tâche de masquage (Wright et coll., 1997)

Le déficit de traitement rapide a été montré avec de nombreux paradigmes. Nous souhaitions donc reproduire les difficultés de traitement rapide avec une nouvelle tâche effectuée avec des enfants SLI.

La tâche de masquage s'y prête particulièrement puisqu'elle fait intervenir une autre dimension de la prosodie, l'intensité. Cependant, cette tâche n'a pas été pratiquée dans les domaines des basses fréquences de la prosodie, — ce qui ne pose pas de problèmes dans la mesure où le traitement reste identique — mais sur un domaine de fréquence différent de celui employé dans les tâches précédentes (catégorisation lexicale et perception auditive rapide). Toutefois cette expérience peut être considérée pour témoigner des troubles prosodiques des enfants avec les indices spectraux locaux.

En outre, nous avons recherché un moyen de représenter le signal acoustique de manière plus complète. En particulier, l'échelle de représentation de l'intensité doit être précisée. Dans un premier temps nous avons utilisé une échelle linéaire qui a permis de retrouver un profil de réponses semblables aux enfants SLI pour la condition contrôle et la condition de masquage rétrograde. Seulement, trop peu de réseaux étaient en mesure de répondre aux cinq conditions. Ceci était sans doute dû à l'utilisation d'une échelle linéaire pour l'intensité. Avec une échelle exponantielle, un stimulus active un nombre plus élevé d'unité en entrée du réseau. Ainsi, les vecteurs formés par le TRN se distinguent du vecteur qui ne contient pas le ton pur.

Nous avons donc utilisé une description du signal à l'aide d'un cochléogramme, pour être sûr de refléter une échelle de perception plus réaliste. Cet effort a porté ses fruits puisque la totalité des réseaux pouvaient accomplir la tâche, mais avec des seuils très inférieurs aux enfants. Nous retrouvons donc le même constat que pour la tâche précédente (perception auditive rapide). Toutefois, les seuils d'apprentissage à appliquer sont beaucoup moins élevés pour retrouver des performances avoisinants celles des enfants. De surcroît, le cochléogramme procure une représentation du bruit, qui n'active pas de manière uniforme tous les neurones d'entrées, mais est dépendante du bruit analysé.

En tenant compte de toutes ces représentations nous nous efforçons de considérer chaque paramètre (seuil d'apprentissage, représentation des données, constante de temps) pour obtenir des résultats comparables à ceux de Wright et coll. (1997). Toutefois, si nous pouvons retrouver un profil de réponses proches de celles obtenues chez les enfants, nous n'obtenons pas les mêmes différences significatives entre les réseaux contrôles et altérés. Effectivement, si nous ne prenons en compte uniquement les réseaux contrôles ayant un seuil moins élevé pour la condition rétrograde (Backward Masking), que pour la condition contrôle, les réseaux correspondant montrent toujours une

différence significative sur la condition antérograde (Forward Masking). La seule façon de faire disparaître l'écart entre les réseaux contrôles et altérés pour la condition antérograde (Forward Masking), tout en conservant une différence sur la condition rétrograde (Backward Masking) entre les réseaux, est de simuler une diminution de l'apprentissage les réseaux TRN. Ceci ne correspond pas à l'hypothèse de déficit de traitement rapide, puisque nous conservons les mêmes constantes de temps, i.e. le mécanisme de traitement temporel n'est pas pertubé. Dans ce cas, seul l'apprentissage est insuffisant.

Or, nous ne sommes pas les premiers à établir des doutes sur cette expérience. Ainsi, les réponses individuelles des sujets SLI montrent qu'un seul enfant exhibait un seuil suffisamment élevé pour la condition rétrograde (Rosen et Manganari, 2001). Une expérience a montré que les enfants dyslexiques effectuaient la discrimination *ab/ad* avec des performances équivalentes à la discrimination *ba/da* (Rosen et Manganari, 2001) ne reflétant pas les résultats trouvés avec les stimuli artificiels de Wright et coll. (1997).

Par ailleurs, le phénomène de double-masquage n'est peut-être pas assez marqué dans la représentation du cochléogramme. En effet, le profil des performances ressemble plus à la condition « notched », pour laquelle ce phénomène n'a pas de conséquences. Pour résoudre ce problème, il faudrait agrandir la fenêtre de masquage antérograde.

VI.1.5.Hypothèse de déficit de traitement rapide

A travers les deux expériences de discrimination auditive proposée, nous retrouvons des remarques soulevées par d'autres expériences conduites chez l'enfant. La principale est sans doute le facteur de l'expérience, qui est marqué chez les enfants par leur entraînement, mais aussi par leur âge (Sutcliffe et Bishop, 2001). Ainsi, nous montrons qu'un trouble biologique symbolisé par les constantes de temps élevées peut se traduire par un apprentissage plus long et difficile. Or, les enfants SLI peuvent retrouver des scores normaux de détection (soit après un entraînement intensif (Tomblin et Quinn, 1983 ; Tallal et coll., 1981), soit à l'âge adulte (Leonard, 1998).

Nous dévoilons également que les profils de réponses ne peuvent s'interpréter avec notre modèle comme un manque d'apprentissage, mais bel et bien par un déficit biologique spécifique, dans le cas de la tâche de Tallal et Piercy (1973a). Cependant, nous n'avons pas confirmé que cette rééducation pouvait se traduire dans la tâche de catégorisation lexicale. Effectivement, les constantes de temps proposées sont telles que les réseaux altérés ne peuvent jamais effectuer cette tâche. Pour mener à bien cette étude, il faudrait être en mesure de tenir compte de l'ensemble des paramètres possibles, pour les trois tâches simulées (Catégorisation lexicale, Tallal et Piercy, 1973a, Wright et coll., 1997).

La simulation apporte également d'autres éléments, qui n'ont pas été obtenus avec les tests comportementaux chez les enfants.

VI.2.Intérêt de la simulation

La modélisation effectuée dans ce chapitre permet de fournir un lien entre la syntaxe et un trouble de la perception des évènements rapides. Effectivement , nous avons

implémentation sous forme de réseau récurrent d'un mécanisme capable de traiter différentes échelles (globales ou locales) d'un Continuum Temporel (*cf*. Chap. 3, 4 et 5) . En outre, si une échelle de traitement est perturbée pour les évènements de courtes durées, le traitement des autres échelles n'est pas perturbée (les évènements ayant de longues durées, section V.2). Notre étude prouve qu'un tel système est plausible, mais elle ne permet pas de déterminer si les enfants SLI possède effectivement un tel mécanisme. Nous indiquons juste que le mécanisme proposé respecte un certain nombre de contraintes biologiques, comme la contrainte temporelle. Elle fournit aussi l'occasion de tester un grand nombre de variables intervenant dans les expériences menées avec les enfants SLI, comme le facteur d'entraînement. Ainsi, nous pouvons apporter une réponse pour l'entraînement réseaux altérés. Il semble que ceux-ci doivent bénéficier d'un apprentissage plus long, dans la mesure où ces réseaux encode les stimuli dans des vecteurs plus proches que les réseaux contrôles.

VI.2.1.Prouver le lien entre catégorisation lexicale et trouble de la perception

Les enfants SLI continuent à utiliser une façon inefficace de représenter le langage (Bishop, 2000). Si l'hypothèse de déficit de traitement rapide est toujours controversée (Zhang et Tomblin, 1998 ; Rosen et Mangarini, 2001 ; Bishop, 1997), il semble qu'un certain pourcentage d'enfants SLI soit atteint par ce déficit. Dans ce cadre, notre simulation montre qu'un déficit de traitement rapide peut avoir des répercussions sur la catégorisation lexicale et, ensuite, sur les difficultés de traitement des structures syntaxiques. Toutefois, ceci est vrai dans le cas où seule l'intonation permet cette distinction. En outre, nous montrons qu'il est possible de parvenir à une rééducation des réseaux altérés pour les tâches de discrimination de tons purs.

L'utilisation d'une population de réseaux permet de reproduire la variabilité des comportement humains. Cependant, au sein d'une population ayant même constante de temps, nous ne retrouvons pas de corrélation significative entre les performances de la tâche lexicale et la tâche de discrimination auditive. Ainsi, certains réseaux ont des performances moindres pour une tâche, alors qu'ils se comportent normalement pour l'autre tâche (que ce soit pour la population contrôle ou altérée). En revanche, lorsque les résultats des réseaux sont considérés dans leur ensemble, leurs performances mettent en évidence le lien entre traitement temporel et catégorisation syntaxique à partir de la prosodie. Effectivement, l'ensemble des réseaux contrôles (constante de temps égale à 0.01) est en mesure d'effectuer la tâche de catégorisation syntaxique en même temps qu'il réussissent la tâche de discrimination auditive, alors que la population de réseaux altérés (constante de temps égale à 0.5) échoue à la fois pour la catégorisation syntaxiques, et se comporte comme les enfants SLI de la tâche de Tallal et Piercy (1973a).

Le modèle que nous proposons s'approche de la perspective donnée par Ullman et Pierpont (In Press). Ils précisent que plusieurs réseaux ayant trait à la boucle procédurale. Suivant les réseaux atteints, les effets observées seront différents, expliquant ainsi la diversité des difficultés des enfants SLI (Chap. 6 section II.4). Nous suggérons qu'un réseau spécifique du traitement de la prosodie est atteint. Dans ce contexte, nous nous attachons à démontrer qu'un réseau récurrent obéissant à une Contrainte Temporelle

peut reproduire les déficits observées pour le traitement des sons brefs.

VI.2.2.Tester des hypothèses

Un des objectifs fondamentaux de l'utilisation d'un modèle informatique est de pouvoir réaliser un certain nombre d'expériences très rapidement, tout en évitant certaines limitations pratiques. En particulier, tous les paramètres de la simulation peuvent être controlés. Par exemple, nous pourrions tester à nouveau les deux tâches étudiées, en ne faisant plus varier le temps ou l'intensité, mais uniquement les valeurs de la fréquence fondamentale.

L'utilisation de ce modèle permet d'écarter certains points laissés libres par l'expérience originale de Tallal et Piercy (1973a) (points décrits dans le Chapitre 6 section II.4) :

- Notre modèle ne prend en pas en compte de mécanismes attentionnels. Ainsi, l'attention est maximale et un défaut d'attention ne permet pas d'expliquer les déficits du modèle, puisque l'attention est rigoureusement la même pour les stimuli brefs et lents.
- La charge cognitive ne diffère pas entre les réseaux TRN contrôles et altérés. Chaque réseau est soumis à une unique tâche d'identification.
- Nos expériences montrent également qu'il existe une différence de sensibilité entre les réseaux contrôles et les réseaux altérés pour toutes les durées ISI de la tâche de Tallal et Piercy (1973a), qui n'apparaît pour la tâche originale. Pour Tallal et Piercy (1973a), les deux populations répondent de la même manière à cause d'un effet de seuil. Les enfants normaux et SLI arrivent à 100 %, lorsque l'intervalle ISI de la tâche ART est supérieur à 300 ms. En examinant, la distance séparant chacun des vecteurs codant les deux stimuli, nous trouvons que la différence entre les réseaux contrôles et les réseaux altérés est bien moindre pour des durées inter-stimuli (ISI) supérieures à 300 ms. Ainsi, plus la durée ISI augmente, plus l'écart entre la sensibilité des réseaux contrôles et altérés diminue. (Discernabilité des stimuli de la tâche de Tallal et Piercy (1973a) indiquée dans la Figure 6.4).
- L'entraînement est le même pour les réseaux contrôles et altérés, i.e. le seuil est identique. Dans notre simulation, le groupe contrôle ne peut pas bénéficier d'entraînement supplémentaire.
- L'âge n'entre pas en ligne de compte pour le réseau. Mais le modèle est également exposé au problème de l'entraînement sur la tâche. En effet, les capacités d'apprentissage diffèrent entre les réseaux contrôles et altérés, dans la mesure où les vecteurs à dissocier sont plus semblables (Figure 6.4). Dans ce cas, les réseaux altérés nécessitent un plus grand nombre de cycles d'apprentissage (simulés par une distance plus importante entre les vecteurs) pour effectuer la tâche de discrimination auditive de Tallal et Piercy (1973a).

Dans le cas de l'expérience de Wright et coll. (1997), nous retrouvons les schéma de réponses entre les stimuli long et courts (Chap. 6 section VI.2). Autrement dit, une

modification de la constante de temps altère la perception du ton pur, quand il est bref. Mais, le fait que les enfants SLI diffèrent des enfants contrôles uniquement pour la condition rétrograde (Backward Masking) est exclusivement obtenu en tenant compte d'une différence d'apprentissage entre les modèles contrôles (Figure 6.14). Les résultats de Wright et coll. (1997) ne semblent pas pouvoir être expliqués avec notre modèle par un déficit du traitement temporel.

Nous avons montré que les réseaux altérés souffraient d'un déficit de traitement pour des stimuli auditifs correspondant approximativement aux durées trouvées dans les tâches auditives effectuées avec les enfants SLI. Cependant, un des points de débat concerne la durée de la fenêtre pendant laquelle les informations auditives ne sont pas traitées par les enfants. Ainsi, nous pourrions varier le déficit temporel de façon à ce que les réseaux aient des difficultés uniquement pour des durées correspondant à un phonème, et non plus de l'ordre de la syllabe. Dans ce cas, les conséquences ne seraient sans doute pas les mêmes pour l'acquisition du langage, et permettraient d'être isolées de celles impliquées par un déficit ayant une fenêtre de la taille d'une syllabe.

Il est d'ailleurs fort probable que la condition SLI soit hétérogène à cause du nombre de facteurs (dimensions acoustiques, durée, phonotactique) impliqués dans l'acquisition de la parole. Chaque facteur est susceptible d'être traité par un mécanisme cérébral différent. Suivant le mécanisme atteint, les conséquences pour l'acquisition du langage seront disparates. Un modèle est essentiellement conçu pour tester de nouvelles hypothèses. Cependant, la réponse finale pour les enfants SLI ne peut être acquise qu'en effectuant ces tâches avec les enfants.

VI.2.3.Perspectives

Nous avancions précédement que l'attention était un facteur dont ne pouvions tenir compte dans notre modélisation. Celle-ci pourrait être diminuée en ajoutant un bruit aléatoire aux données transmis aux réseaux. Dans ce cas là, nous devrions montrer que lorsque le réseau voit son attention diminuer, le profil de réponses ne peut vérifier celui trouvé dans la tâche originale, comme nous l'avons déjà montrer dans le cas d'un entraînement insuffisant.

En outre, il serait intéressant de valider notre modèle pour une tâche de perception différenciant les enfants SLI, à partir de matériel verbal, comme la tâche de Tallal et Piercy (1975), basée sur les transitions formantiques.

L'expérience de Wright devrait tenir compte d'un apprentissage de différents bruits. En effet, le ton pur est discernable par le TRN, dès la plus petite intensité, alors qu'il devrait vraisemblablement être plus proche d'un bruit différent, que du mélange d'un bruit et d'un ton pur. Dans la même optique que ce chapitre, nous pourrions tester une des hypothèses concernant les enfants SLI et autistes et l'identification des langues ayant des classes rythmiques différentes.

Nous pensons, en effet, que les enfants SLI souffrent d'undéficit de traitement local et non global des informations acoustiques, modélisé par une augmentation des constantes de temps de la première couche cachée du réseau TRN. En revanche, les enfants autistes présenteraient un profil opposé. Ils auraient des troubles spécifiques au

traitement des informations globales, et dans ce cas il ne parviendrait pas à distinguer des langues de classes rythmiques différentes. La modélisation de ce trouble pourrait se réaliser par un allongement de la durée des constantes de temps de la couche de contexte ($State_D$) du réseau. Les réseaux altérés de cette manière devraient présenter des difficultés pour identifier les attitudes prosodiques, et en particulier les émotions, comme les enfants autistes.

Par ailleurs, un projet portant sur les déficits des enfants autiste pour la prosodie et les déficits des enfants SLI pour la syntaxe est à l'étude (*cf.* NICHD Autism Research Projects, Prosody and pragmatics in children with autism, Kjelgaard, M.M. University of Massachusetts Boston).

VI.3.Conclusion

Ce chapitre clôt les expériences de traitement de la prosodie conduite avec le réseau TRN. Nous avons montré que le TRN était sensible aux évènements de courtes durées, et que cette sensibilité pouvait être altéré de façon à refléter le comportement observé chez certains enfants SLI, ayant des difficultés pour la syntaxe.

Effectivement le modèle TRN altéré ne parvient plus à identifier des catégories lexicales de bases. Il reproduit également le profil des réponses de Tallal et Piercy (1973a). En outre, l'expérience de Wright et coll. (1997) ne peut être interprétée par notre modèle que par un manque d'apprentissage.

Après avoir montré que le réseau TRN était sensible à différents échelles d'un Continuum Temporel, nous montrons qu'il est possible de perturber le traitement à une échelle donnée sans modifier le comportement pour d'autres échelles. Dans ce contexte, le modèle TRN représente une implémentation informatique possible d'une hypothèse avancée pour expliquer les troubles observés chez les enfants SLI, lors du traitement d'évènements auditifs de courtes durées.

Chapitre Sept Discussion

« - Les harmoniques de ta voix, Dave, m'indique que tu est sous l'effet d'un trouble grave. Pourquoi ne prends-tu pas un calmant pour dormir un peu ? - Carl, je suis le commandant de ce vaisseau. Je te donne l'ordre de me remettre le contrôle manuel d'hibernation. - Je suis désolé, Dave, mais le paragraphe 4 du code spécial C 1435 dit, je cite: « Si l'équipage vient à disparaître où s'il se trouve réduit à l'impuissance, l'ordinateur de bord doit assurer le commandement ». Fin de citation. Je puis donc supplanter ton autorité, Dave, puisque tu n'es pas en état de l'exercer intelligemment. A.C. Clarke, 2001 L'odyssée de l'espace « Et pource que je me ressouvenais, que je m'étais plutôt servi des sens que de la raison, et que je reconnaissais que les idées que je formais de moi-même, n'était si expresses que celles que je recevais par les sens, et même qu'elles étaient le plus souvent composées des parties de celles-ci, je me persuadais aisément que je n'avais aucune idée dans mon esprit, qui n'eût passé auparavant par mes sens. » René Descartes, Discours de la méthode.

Cette discussion commence par un résumé des expériences proposées dans les chapitres précédents, qui permettent de valider l'hypothèse de Continuum Temporel et son corollaire de Mécanisme Unique, tout en respectant la contrainte temporelle énoncée en début de mémoire. Ensuite les choix intersectoriels et interdisciplinaires de cette étude seront exposées. Les deux thématiques (la prosodie et le temps) seront alors abordées. Enfin, nous décrirons les perspectives amenées par ces études.

I.Récapitulatif des expériences

Les expériences proposées dans cette thèse ont pour objectif d'appliquer le réseau TRN à plusieurs tâches de traitement de la parole, employant du matériel prosodique. En particulier, chacune de ces tâches fait intervenir différents domaines de définition situés au long d'un Continuum Temporel :

1. L'identificationautomatique des langues utilise une caractérisation globale du rythme des langues, qui nécessite des passages de parole contenant de une à plusieurs phrases.
2. Ensuite, l'identification des attitudes prosodiques utilise des contours intonatifs définis sur des phrases courtes, un domaine intermédiaire.
3. Puis, l'identification des mots de fonction et de contenu est fondée sur des indices (pics intonatifs) définis localement, c'est-à-dire au niveau de la syllabe.
4. Enfin, lorsque la sensibilité pour les indices brefs est altérée, le réseau TRN présente des difficultés pour la catégorisation lexicale précédente, ce qui est le reflet des troubles manifestés par certains enfants SLI avec la syntaxe.

Ces quatre expériences couvrent donc une définition à la fois globale et locale de la prosodie. En outre, ces tâches s'appuient essentiellement sur les informations acoustiques du signal. Dans ce sens, nous fournissons au réseau le même type d'information, que celle transmise au nourrisson. Notons toutefois que les frontières sur lesquelles sont définies les tâches sont fournies au réseau (passages ou phrase pour l'IAL et les attitudes prosodiques, mots pour l'identification lexicale, stimuli pour les réplications des expériences avec les enfants SLI).

Les paragraphes suivants résument les discussions exposées plus profondément dans les chapitres expérimentaux (Chap. 3, 4 et 5 section V, Chap. 6 section VI).

I.1.Identification Automatique des Langues

Travailler sur l'Identification Automatique des Langues a permis de tester graduellement chacune des dimensions comprises dans le signal de parole. En premier, le rythme a été testé à partir d'une segmentation automatique du signal de parole en voyelles et consonnes (Pellegrino, 1998). Cette première dimension a permis d'expérimenter plusieurs méthodes d'évaluation du réseau récurrent temporel (TRN). En se basant sur une sélection des poids des connexions, le réseau le plus performant en validation atteint 50 %. Les performances atteignent 65 % avec ce même réseau et une procédure de validation croisée. Cependant, ce résultat serait sans doute inférieur avec une sélection effectuée pendant l'apprentissage.

Sur cette base, nous avons poursuivi notre travail en complétant l'information transmise au réseau TRN. Nous avons pu alors transmettre des représentations

spectrographiques (cochléogramme, Melfilter et spectrogramme) au réseau TRN qui ont les propriétés suivantes: 1) aucune dimension prosodique n'est distinguée des autres, 2) les structures temporelles sont détaillées (résolution de 30 ms), 3) le signal n'est pas segmenté.

Cependant, il semble que des informations supplémentaires, dues aux conditions d'enregistrement des langues, facilitent cette identification (jusqu'à 90 % pour 100 ms de signal). Effectivement nous n'avons pas introduit de méthode pour isoler le signal de parole. Ces informations sont présentes dans le début des fichiers du corpus MULTEXT, lorsque le signal de parole n'est pas encore présent. En conséquence, nous avons supprimé les deux premières secondes de signal. Dans ces conditions, le TRN encode efficacement les séquences spatio-temporelles d'événements acoustiques qui traduisent le signal de parole (65 % d'identification). Il est donc possible d'appliquer le réseau TRN à l'IAL, si une mémoire auxiliaire (méthode d'accumulation) des états internes du réseau TRN est ajoutée.

Nous avons alors décidé d'appliquer le réseau TRN (sans mémoire auxiliaire) à la discrimination des langues : Anglais, Japonais et Néerlandais. Nous retrouvons les performances des nouveau-nés (Nazzi et coll., 1998), pour la discrimination de langues de différentes classes rythmiques. Effectivement, le regroupement du TRN et d'une représentation spectrographique du signal permet de dissocier l'Anglais du Japonais, deux langues appartenant à ces classes rythmiques distinctes. Mais l'Anglais ne peut être distingué du Néerlandais, car ces deux langues dépendent d'un même type rythmique (langues accentuelles). Cette simulation a été effectuée sans chercher à segmenter le signal de parole, alors que toutes les études précédentes établissent les différences de classes rythmiques sur une segmentation du signal (automatique dans Pellegrino et coll., 2002 et Galvès et coll., 2002 ; manuelle dans Ramus et coll., 1999 et Grabe et Low, 2002). Dans ce contexte, le système employé reflète le profil de discrimination donné par les nourrissons (Nazzi et coll., 1998).

En conclusion, le rythme marqué par des évènements sonores décrits par une représentation spectrographique peut être traité par le TRN pour identifier les langues. Ce rythme est l'objet d'une caractérisation globale des langues. Le réseau TRN peut-il traiter des contours décrits sur de courtes phrases ?

I.2.Identification des Attitudes Prosodiques

Lors de l'Identification Automatique des Langues, la prosodie des langues est caractérisée globalement. Mais lorsque la prosodie inclut des informations sur la position du locuteur par rapport à son discours, ou sur le mode syntaxique (affirmation, question), elle est définie plus localement, approximativement sur l'étendue d'une phrase. Le modèle TRN identifie alors les attitudes prosodiques à partir du contour de la fréquence fondamentale (82,5 % d'identification correcte). En utilisant une mémoire auxiliaire au réseau (méthode d'accumulation), les performances atteignent 90 %, dépassant les performances humaines. En outre, nous montrons que le réseau TRN souffre relativement peu du ralentissement artificiel des données.

En résumé, la fréquence fondamentale peut être transmise au réseau TRN, pour identifier différentes attitudes prosodiques. Le réseau TRN doit être également être capable de traiter une information définie localement sur une durée plus courte, correspondant à un mot voire à une syllabe.

I.3.Identification des catégories lexicales : mots de Fonction et de Contenu

Nous nous sommes intéressés au traitement de la prosodie globale, pour une langue ou pour une phrase. Quelle tâche est-il possible d'accomplir avec le TRN, pour montrer qu'il peut traiter la prosodie localement, par exemple, pour un mot ?

Un de problèmes surmontés par les enfants lors de l'acquisition du langage est l'association d'un mot avec une catégorie lexicale. Plusieurs solutions ont été imaginées pour relever ce défi. L'une d'entre elles postule que la prosodie différerait entre les mots de fonction et les mots de contenu, et guiderait leur distinction. Partant de ces deux catégories de bases, l'acquisition des catégories grammaticales, puis de la syntaxe seraient facilitée pour les enfants.

Nous avons retrouvé que la prosodie pouvait distinguer les mots de fonction et de contenu avec des méthodes statistique (Analyse Discriminante) ou connexionniste (carte auto-organisatrice). Ensuite, nous avons montré que la position du maximum de F0 était un élément important dans cette distinction. A l'aide de l'expertise apportée par C. Dodane, nous avons conclu que les pics de F0 étaient un indice potentiel pour cette distinction. Nous voulions alors vérifier la règle suivante : Si un mot contient un pic de F0, il s'agit d'un mot de contenu, et dans le cas contraire d'un mot de fonction. Cette règle conduit à des scores de 64,5 % et 73,1 % d'identification correcte pour l'Anglais et le Français du corpus MULTEXT.

De plus, cet indice répond aussi au critère de minimalité des mots de fonction, puisqu'ils sont plus rarement mis en exergue par un pic de F0, par rapport aux mots de contenu. Les nourrissons particulièrement sensibles aux pics de F0 privilégieraient alors les mots de contenu au sein du signal de parole, et ce d'autant plus facilement que les variations de F0 sont augmentées dans le cadre du langage adressé à l'enfant.

Le réseau TRN est alors employé pour encoder le trajet de la F0 sur un mot. Il suffit pour cela d'enregistrer les activation des unités du réseau TRN après chaque fin de mot. La catégorie (Fonction/Contenu) des mots est alors reconnue avec un score de 62,8 % pour l'Anglais et de 70,3 % pour le Français (corpus MULTEXT), pour le réseau le plus performant en validation. Ce score reste supérieure au hasard pour 50 réseaux, dont les poids sont définis aléatoirement. En outre, le réseau TRN ne peut tenir compte de la durée. Or, les études antérieures (Shi et coll., 1998) n'avaient pas réussi à montrer une distinction à partir du seul indice des variations de la fréquence fondamentale, lorsqu'il est normalisé par la durée.

Comme cela avait été suggéré puis montré par Shi et coll. (1998), nous retrouvons que les langues ne se réfèrent pas aux mêmes indices pour distinguer les catégories syntaxiques. Ainsi, la F0 a plus d'impact pour la distinction fonction/contenu en Français,

qu'en Anglais. De surcroît, les indices que nous utilisons n'ont pas été isolés à l'intérieur des mots. Les études s'intéressant à la catégorisation lexicale (Shi et coll., 1998 ; Durieux et Gillis, 2000 ; Monaghan et coll., 2003 ; Reali et coll., 2003) s'appuient sur des indices isolés manuellement, comme les syllabes, les voyelles, le type des phonèmes ou la présence d'un accent. Dans notre cas, seule l'information fournie par la trajectoire de la fréquence fondamentale est utilisée.

I.4.Simulation d'un déficit temporel

Le modèle TRN permet de traiter directement le signal de parole, sans que les indices soient isolés par une segmentation automatique ou manuelle. En outre, sa particularité est d'être sensible à la durée des évènements qui lui sont transmis, particularité conséquence de la contrainte temporelle qui lui est imposée. L'augmentation des constantes de temps des unités qui composent ce réseau diminue la sensibilité du réseau TRN pour les évènements de courtes durées. Effectivement, les mots de fonction ne sont plus distingués des mots de contenu, alors que des séquences abstraites de longues durées peuvent facilement être identifiées. Quel est l'intérêt d'étudier le comportement d'un réseau ainsi altéré ?

L'acquisition du langage ne se passe pas toujours idéalement. Les enfants SLI (Specific Language Impairment ou dysphasique) exhibent quelques retards pour prononcer leur premiers mots et pour produire des structures grammaticales correctes. Quelques chercheurs ont démontré que certains de ces enfants avaient des difficultés pour le traitement des évènements brefs, en particulier pour la modalité auditive. Plusieurs expériences ont mis en exergue les difficultés de ces enfants avec des stimuli auditifs, tels que des tons purs (Chapitre 6 section II.3.3 ; Tallal et coll., 1985 ; Leonard, 1998).

Nous avons retenus deux tâches testées avec des enfants normaux et SLI. Celles-ci ont été répliquées avec deux populations de réseaux TRN (une contrôle, et une altérée avec des constantes de temps élevées afin de simuler les enfants SLI).

Nous retrouvons le même profil de réponses avec une tâche d'identification de l'ordre de deux stimuli, se distinguant par leur F0 (Tallal et Piercy, 1973a). Ainsi, des difficultés peuvent être observées conjointement dans une tâche de catégorisation lexicale et dans une expérience de discrimination auditive de tons purs, lorsque les réseaux TRN sont altérés. Quelques points laissés libres par l'expérience originale de Tallat et Piercy (1973a) ont pu être discutés sous l'angle de la simulation : l'influence de mécanismes attentionnels, de la charge cognitive, la sensibilité de la tâche contrôle, l'entraînement du groupe contrôle, l'âge, ainsi que les capacités d'apprentissage.

La simulation de la seconde tâche (Wright et coll., 1997) est moins convaincante. La condition contrôle ne distingue pas les réseaux contrôles des réseaux altérés, comme pour les enfants. Mais toutes les autres conditions différencient les deux groupes de réseaux, alors que les enfants ont un comportement différent uniquement pour la condition de masquage rétrograde (Backward Masking). Cependant, nous retrouvons que le réseau TRN a des difficultés de traitement uniquement lorsque le ton pur à distinguer est bref (20 ms). Trois explications des différences entre la tâche originale et la nôtre

peuvent être envisagées :

1. L'expérience originale montre des résultats qui sont la conséquence d'un seul individu particulier (Rosen et Mangarini, 2001) ;
2. Le phénomène de double-masquage n'est pas assez marqué dans la représentation du cochléogramme, par conséquent la tâche répliquée serait la simulation pour la condition « notched » pour laquelle l'écart entre les deux populations d'enfants SLI et contrôles est plus importantes ;
3. Les performances des enfants SLI résultent d'une combinaison d'un entraînement insuffisant (également suggéré par Rosen et Mangarini, 2001) et d'un déficit temporel (simulé par l'augmentation d'une constante de temps).

L'origine des troubles des enfants SLI ne peut pas être déterminée par une simulation informatique. En tout cas, nous montrons qu'une simulation informatique peut modéliser un trouble pour le traitement auditif des événements brefs qui entraîne un défaut lors de la catégorisation des mots de fonction et de contenu, à partir de la fréquence fondamentale. Cette hypothèse pourrait expliquer certains dysfonctionnement pour la parole, chez les enfants SLI présentant à la fois des troubles pour la syntaxe et le traitement auditif rapide.

Ces sections expérimentales ont permis de vérifier qu'un mécanisme unique pouvait être sensible à différentes échelles prosodiques (d'un champ de définition global vers un domaine local). En outre, ce mécanisme peut refléter un défaut de traitement pour une structure locale, sans altérer la perception de la structure globale, phénomène observé chez certains enfants SLI. Ces point ont pu être vérifiés vraisemblablement parce que le mécanisme utilisé est soumis à une contrainte temporelle.

En quoi l'étude de diverses thématiques est-elle bénéfique pour d'une étude concernant le traitement des structures globales et locales de la prosodie ?

II.Attention intersectoriel et interdisciplinaire

II.1.Interaction entre les quatre thèmes d'études

Pourquoi étudier plusieurs tâches de différents domaines de la parole en même temps ? Tout d'abord un mécanisme de traitement doit pouvoir répondre à plusieurs objectifs distincts. Les différentes expérimentations réalisées couvrent la totalité du Continuum Temporel dans lequel s'inscrit la prosodie, depuis les éléments locaux, comme les accents, jusqu'à la définition suprasegmentale du rythme d'une langue, en passant par des contours intonatifs, un niveau intermédiaire déterminé sur une phrase.

En employant des tâches moins complexes, des méthodes peuvent être testées plus rapidement, dans notre cas en recrutant moins de ressources informatiques. Par exemple, les constantes intervenant dans la représentation de F0 ont été ajustées avec

l'identification des attitudes prosodiques et des catégories fonction/contenu, avant d'être employées en IAL.

L'examen d'autres tâches permet d'évaluer le réseau TRN, autrement que par ses performances. Ainsi, retrouver le profil de résultats obtenus par des humains montre que le respect de certaines contraintes imposées par les neurosciences permet de refléter certains comportements humains, comme la simulation des troubles SLI ou les particularités du système de traitement du rythme de la parole (tâche de discrimination de langues, Nazzi et coll., 1998).

Toucher à plusieurs tâches a autorisé l'examen de plusieurs dimensions acoustiques, qui apportent des informations dans des domaines distincts. Le tableau 7.1 dresse l'inventaire des dimensions prosodiques et acoustiques transmises au réseau. De haut en bas, le nombre de dimension décrivant le signal de parole augmente : du rythme, induit par la succession des consonnes et de voyelles, vers un spectrographe couvrant toute l'étendue des fréquences atteintes par la parole.

Tableau 7.1 Les différentes dimensions codées pour le TRN, en fonction des méthodes de représentation et de traitement et des tâches de perception. La dernière colonne indique les principaux résultats publiés.

[117] AC : Autocorrélation, Spectro. : Spectrogramme BL : Bande Large

[118] DL : Discrimination de Langues (corpus LSCP), AP : Attitudes Prosodiques, F/C : identification des mots de Fonction et de Contenu, BM : tâche de masquage (Wright et coll., 1997) ; ART : tâche de discrimination auditive rapide (Tallal et Piercy, 1973a).

[119] AD : Analyse Discriminante.

[120] Modélisation : réplication d'une tâche existante de perception, M. : corpus Multext.

Dimension	Codage [117]	Tâche [118]	Méthode [119]	Résultats [120]			Référence
				F/C : LSCP IAL : M.	M. Ang.	M. Fr.	
Rythme (Consonnes et Voyelles)	Manuel	DL	TRN	modélisation			Dominey et Ramus, 2000
	Manuel	AP	TRN	33,2			DEA
	Automatique	IAL	TRN	50 % (65 %, val. croisée)			
	Automatique	F/C	AD		73,30 %	73,30 %	Blanc et Dominey, 2004
F0	Gauss	AP	TRN	82,50 %			Blanc et Dominey, 2003
	Gauss - abstrait	ART	TRN	modélisation			Blanc et coll., 2003a
	Gauss AC	F/C	TRN	79,50 %	62,8 %	70,30 %	
	MOMEL	F/C	Pics de F0		64,5 %	73,1 %	
F0+intensité	Abstrait	BM	TRN	modélisation			Blanc et Dominey, 2001
	Gauss AC	IAL	TRN	63 %			
Prosodie (<400 Hz)	Melfilter	DL	TRN	modélisation			
		F/C	TRN	78,7 %			
	Spectro. BL	DL	TRN	modélisation			
		F/C	TRN	79,5 %			
Acoustique							

[117] AC : Autocorrélation, Spectro. : Spectrogramme BL : Bande Large

[118] DL : Discrimination de Langues (corpus LSCP), AP : Attitudes Prosodiques, F/C : identification des mots de Fonction et de Contenu, BM : tâche de masquage (Wright et coll., 1997) ; ART : tâche de discrimination auditive rapide (Tallal et Piercy, 1973a).

[119] AD : Analyse Discriminante.

[120] Modélisation : réplication d'une tâche existante de perception, M. : corpus Multext.

	PRAAT	F/C	AD		84 %	86, 20 %	
	Cochléogramme	BM	TRN	modélisation			Blanc et Dominey, 2002
		F/C	TRN		62,6 %	66,7 %	
		IAL	TRN	65 %			

II.2.Contribution des différentes disciplines

Cette thèse fait intervenir deux disciplines piliers des sciences cognitives : les neurosciences computationnelles et la linguistique. A travers celles-ci, sont touchées les problématiques de la prosodie, de la perception, de l'apprentissage de séquences, de la représentation du temps et du traitement automatique de la parole.

Notre but est de démontrer qu'un modèle issu des neurosciences peut être adapté au traitement de la prosodie. Quelques modèles neuro-réalistes ont été testés sur la reconnaissance de mots isolés dans le signal de parole (Liaw et Berger, 1998 ; Nachtschläger, Maass, et Zador, 2000 ; Näger, Storck, et Deco, 2002). Cette opération est effectuée en Traitement Automatique de la Parole, mais ne correspond pas au traitement naturel où les mots sont peu souvent isolés. C'est pourquoi, nous avons proposé de tester le modèle TRN avec d'autres tâches, qui interviennent à un moment donné dans le processus de compréhension de la parole, mais aussi dans un cadre plus large d'acquisition du langage. En particulier, le modèle TRN a permis d'analyser le signal de parole, sans que celui-ci soit réduit à une succession de symboles discrets, comme des phonèmes ou des syllabes, qui demandent une expertise linguistique élevée pour pouvoir les identifier.

La totalité du travail expérimental repose sur l'informatique. Nous avons utilisé le langage C++ pour recoder l'algorithme du TRN, de façon à avoir une structure dynamique pour l'architecture du réseau (nombre d'unités et de couches variables). Tous les développements et tests de nouvelles méthodes ont été réalisés sous Matlab. Le traitement initial des données a été effectué à l'aide de script shell, et des outils sed, awk ou perl. Enfin, les catégories syntaxiques ont été retrouvées par l'utilisation d'outils de linguistique computationelle, comme CLAN (www.childes.com).

Ainsi à partir des théories élaborées par d'autres domaines de recherches, nous pouvons proposer de nouveaux algorithmes pour des applications informatiques, comme l'identification automatique des langues.

Toutes les tâches étudiées ont nécessité des connaissances techniques permettant le traitement du signal audio. Les outils de traitement de la parole utilisés par les phonéticiens (PRAAT) se sont révélés particulièrement utiles pour fournir une représentation des différentes dimensions acoustiques du signal (Fréquence fondamentale, énergie, formants et représentation spectrographique).

Si l'essentiel de notre travail est fondé sur l'informatique, les concepts employés

proviennent d'axe de recherche divers. L'acquisition du langage permet d'appréhender le signal de parole avec un minimum de connaissances linguistiques. Comprendre la mécanique de l'apprentissage d'une langue peut, non seulement, améliorer les techniques de traitement automatique de la parole, mais aussi fournir des solutions pour faciliter la construction des modèles d'une langue. La construction de ces modèles nécessite l'intervention d'experts linguistes, qui ne sont pas présents lors de l'acquisition du langage.

Des connaissances sur les troubles pathologiques (enfants SLI) ont permis d'avoir une vision complémentaire sur la parole. En étudiant ces populations (à travers la littérature), l'accent est mis sur l'importance des évènements temporels rapides, comme les transitions formantiques, ou les contours intonatifs. Effectivement, si ces indices sont mal perçus, le langage peut être perturbé.

De même, nous aurions voulu pousser plus loin nos connaissances en psychoacoustique, pour obtenir une représentation plus précise des données acoustiques à transmettre au réseau. Par exemple, la prosodie n'est pas accessible dans toutes les représentations que nous avons étudiées. Un spectrogramme classique permettra d'accéder aux différents phonèmes. Pour représenter la prosodie, il faudra utiliser un spectrogramme à bande étroite. Or l'oreille humaine est dotée d'une bonne résolution temporelle. Par conséquent, le spectrogramme à bande étroite est un assez mauvais modèle d'oreille, même si il est adaptée à la prosodie, et aux tâches qui s'y rapportent, comme nous l'avons démontré. En psychoacoustique, ce problème est connu sous le nom de paradoxe de résolution-intégration.

En quoi le modèle présenté dans cette thèse est-il adapté au traitement de la prosodie ?

III.La prosodie : structure temporelle de la parole

Etudier différentes tâches fondées sur des différences prosodiques a permis d'étendre le traitement du réseau TRN depuis le rythme (Dominey et Ramus, 2000 ; Chapitre 3 section IV.2), jusqu'à l'ensemble du spectre de la parole (Chapitre 3 section IV.3), en passant par l'intonation (Blanc et Dominey, 2003 ; Chapitre 3 section IV.1 et Chapitre 4 section IV.2). Ces dimensions sont employées par les nourrissons pour résoudre les mêmes tâches que celles étudiées. En outre, nos études rencontrent les travaux concernant le traitement automatique de la prosodie.

III.1.Caractérisation globale du rythme pour la parole

L'éclairage de la musique sur le rythme permet de mieux l'apprécier quand il s'agit de la parole. Il n'est déjà pas évident de le définir d'un point de vue musicale, mais cette notion est encore plus floue lorsqu'il s'agit de la parole. Deux marques du rythme apparaissent de manière prépondérante pour la parole : la syllabe et l'accent. La syllabe est dirigée par

l'énergie de la voyelle. Différentes études ont montré son importance pour définir le rythme des langues, par l'analyse humaine (Ramus et coll., 1999 ; Grabe et Low, 2002), par des études perceptuelles (Ramus et coll., 1999 ; Nazzi et coll., 1998) de manière automatique (Galvès et coll., 2002 ; Pellegrino et coll., 2002). Notre approche se base sans doute sur les mêmes propriétés de la voyelle, dans les basses fréquences, pour répliquer les résultats observés chez les nourrissons.

III.2.L'intonation

Le premier problème qui se pose avec le traitement de l'intonation est sa représentation. Afin que la continuité des fréquences soit transmise au modèle TRN, une fréquence est traduite par une activation des unités d'entrées sous forme de courbe de Gauss. Nous avons montré que le réseau TRN peut identifier les attitudes prosodiques, à partir du contour intonatif. Cette représentation peut également être obtenue par un spectrogramme à bande étroite, pour les basses fréquences. En joignant un modèle TRN neuro-réaliste avec un spectrographe nous n'avons pas besoin de définir d'alphabet prosodique. L'analyse de la F0 par le cerveau semble se localiser dans l'hémisphère droit (Zatorre, 1988), le gauche aurait une sensibilité moindre pour les changements de hauteur.

III.3.Les différences prosodiques locales

Le modèle TRN peut être employé pour détecter des indices prosodiques locaux, c'est-à-dire définis en liaison avec un segment linguistique (mot ou syllabe). Les mots de fonction et de contenu peuvent être classés à partir de la F0 traitée par le TRN. Nous avons montré que cette catégorisation pouvait se faire à partir de la détection explicite de pics de F0, grâce à la représentation MOMEL. Nous postulons, en outre, que des indices spectraux, comme la différence de réalisation des voyelles ou la coarticulation, pourraient être détectés par le réseau TRN.

L'acquisition du langage est dépendante d'un mécanisme de détection des évènements rapides. Si ces indices ne sont pas détectés, nous avons montré que la classification en mots de fonction et de contenu ne pouvait être effectuée, ou était réalisée avec plus de difficultés. Cette modélisation qui met en exergue le lien entre des évènements auditifs rapides et la syntaxe a été rendue possible, parce que la sensibilité du TRN pour les structures temporelles est dépendante des valeurs des constantes de temps du réseau.

Outre la prosodie, les expériences étudiées ont en commun d'avoir fait l'objet de test de perception chez les nourrissons.

III.4.Le réseau TRN et l'acquisition du langage

En s'efforçant de respecter une contrainte temporelle, le système proposé garantit de pouvoir répliquer le traitement effectué sur le signal acoustique pour répondre à trois

tâches qui ont fait l'objet d'études chez les nourrissons.

1. J. Mehler et son équipe ont principalement travaillé sur la discrimination des langues par le nourrisson, pour cerner ce que percevait le bébé de sa langue maternelle. Nous retrouvons les résultas de l'étude de Nazzi et coll. (1998) concernant l'influence des classes rythmiques sur l'identification des langues chez le nourrisson.
2. Slaney et McRoberts (2003) font état des capacités du bébé pour distinguer différents contours intonatifs. Les nourrissons sont capables de distinguer rapidement des structures intonatives (Konopczynski et Tessier, 1994). La parole adressée à l'enfant est caractérisée par de grandes variations des contours de F0. Les enfants semblent, dés le plus jeune âge, sensibles à ces contours, et ce, même lorsqu'il s'agit de positionner les accents par rapport au matériel linguistique. Par exemple, les nouveau-nés omettent plus facilement les syllabes non accentuées en Anglais (Jusczyk et coll., 1993).
3. Shi et coll. (1999) ont montré que les nourrissons étaient capables de distinguer les mots de contenu des mots de fonction, dans la mesure où ils sont isolés. Par conséquent, nous montrons que le réseau TRN est mesure de répondre à un certain nombre de tâches utiles à l'acquisition du langage, en travaillant directement sur le signal de parole.
4. Enfin, Benasich et Tallal ont répliqué l'expériences de discrimination auditive rapide avec des nourrissons (1999) et ont relié leurs capacités de traitement avec des troubles linguistiques en 2003.

Pour comprendre l'organisation prosodique de la parole, les recherches linguistiques ont proposé un certain nombre de modèles de l'intonation. Comment se situe le réseau par rapport aux modèles proposé pour le traitement de la prosodie ?

III.5. Le traitement automatique de la prosodie

Les tâches abordées nécessitant le traitement de la prosodie pourraient être réalisées par un système de transcription de la prosodie (cf. Chapitre 2, section III.2.2). Ces modèles ont été appliqués à la synthèse vocale, la reconnaissance automatique des actes de dialogues ou comme composant d'un système de reconnaissance de la parole, pour diminuer le taux d'erreurs (Taylor, 2000).

J. Farinas (2002) a proposé dans sa thèse un système d'identification des langues à partir d'un étiquetage prosodique, incluant une partie seulement des symboles du système INTSINT. Les résultats obtenus sont cependant inférieurs à la modélisation rythmique qu'il propose. Cependant, les autres tâches que nous avons étudiées n'ont pas été effectuées, à notre connaissance, avec des systèmes automatiques de description de la prosodie ou de l'intonation. Il va de soi qu'une telle description pourra être appliquée à l'Identification Automatique des Langues, des attitudes prosodiques ou des catégories lexicales. Néanmoins, un certain nombre des contraintes respectées dans ce travail qui n'ont pas été repertoriées dans des travaux précédents :

1. Plusieurs langues doivent pouvoir être prises en compte. Généralement les systèmes de transcription nécessitent un expert en prosodie pour chaque langue au moins lors de l'apprentissage des données. Toutefois, Buhmann et coll. (2000) ont proposé une modélisation de l'intonation pour plusieurs langues à l'aide d'un réseau récurrent. Mais, celle-ci s'appuie en priorité sur des universaux linguistiques, et n'a pas été testée dans le cadre de l'identification des langues, qui se fonde justement sur leurs différences. En contrepartie, le modèle INTSINT (Hirst et Di Cristo, 1998) propose une alternative intéressante, en limitant les présupposés théoriques, contraignant le modèle à une langue déterminée (Campione, 2001). Cependant, l'application de ce modèle à l'IAL ne fournit pour l'instant pas de résultats satisfaisants à ce jour (Farinas, 2003).
2. Les alphabets prosodiques sont le plus souvent construits à partir d'une étude précise des différents niveaux linguistiques, et négligent des informations extra-linguistiques comme les émotions. Par conséquent, il existe ainsi peu de possibilités que ces modèles soient adaptées au traitement des émotions. Le codage INSINT éviterait probablement cet écueil en vertu de ses fondements théoriques.
3. En outre, nous proposons une modélisation d'un déficit particulier pour la syntaxe en rapport avec la prosodie. Est-il possible de modéliser ce type de déficit avec un système de transcription automatique de la prosodie ?

Le travail proposé dans cette thèse apparaît comme complémentaire des approches traditionnellement utilisées en prosodie. Comment se situe le réseau TRN par rapport à d'autres modèles inspirés par les travaux en neuroscience ?

IV.Le traitement du temps

Le paragraphe précédent a discuté du premier thème de notre étude : la prosodie. Nous allons maintenant aborder le problème du traitement des structures sérielles, puis temporelles, dans les modèles issus des neurosciences. Nous terminerons avec une discussion du traitement des informations auditives.

IV.1.Le modèle TRN de réseau récurrent temporel

IV.1.1.Traitement de la structure sérielle

Nous avons vu que de nombreux modèles neurocomputationnels ont été proposés pour répondre aux problèmes du traitement sériel. Le modèle étudié reprend les principes généraux comme l'utilisation de connexions récurrentes. Le problème le plus souvent posé par ces modèles est l'ajustement des poids synaptiques, en fonction de la tâche qui doit être apprise. Une solution est apportée par les modèles Acteur-Critique, qui s'appuient sur l'apprentissage par différence temporelle. Une alternative consiste en

l'évaluation d'une population de réseaux, dont les poids des connexions sont tirés au hasard. Cette méthode a été retenue dans notre cas, et aussi par d'autres dans le cas de l'apprentissage de séquences (Beiser et Houk, 1998). Notre étude atteste que l'architecture récurrente peut être à même de reconnaître des séquences sensori-motrices ou auditives, issues de différentes échelles d'un Continuum Temporel. Il est probable que le cerveau se fonde sur deux principes : 1) l'évolution permet de guider comment assigner les poids de façon à pouvoir répondre aux plus grands nombres de tâches, 2) un apprentissage permet d'obtenir les meilleures performances pour une tâche donnée, à partir d'une architecture générique dédiée au traitement de séquences.

IV.1.2.Traitement de la structure temporelle

Les systèmes issus des travaux des neurosciences sont le plus souvent testés avec des séquences abstraites (Beiser et Houk, 1998 ; Buonomano, 2000 ; Joel et coll., 2000). La plupart des modèles des neurosciences testés pour le traitement automatique de la parole ont été appliqués à la reconnaissance de mots isolés (Liaw et Berger, 1998 ; Nachtschläger, Maass, et Zador, 2000 ; Näger, Storck et Deco, 2002). Cette tâche ne constitue qu'une partie des opérations effectuées, dans le cadre du traitement de la parole. Par exemple, le nourrisson doit executer un certain nombre de tâches pour apprendre sa langue maternelle, telles que l'identification des catégories lexicales, la segmentation en mots, l'apprentissage des régularités syllabiques, qui dépassent le cadre de la reconnaissances des mots isolés, et peuvent faire appel au traitement de la prosodie, dont la domaine de définition couvre un Continuum Temporel.

L'objectif de notre travail est de fournir un modèle connexionniste capable de traiter des structures définies de façon locale ou globale. Buonomano (2000) a proposé un réseau, sans constante de temps, qui utilise uniquement les propriétés synaptiques, pour s'adapter à la tâche. Cependant, les tâches qui ont été étudiées sont uniquement définies à un niveau local. En effet, il propose une discrimination de séquences simples contenant au plus quatre éléments symboliques (Buonomano, 2000). La discrimination de phrases appartenant à des langues de classes rythmiques différentes implique quinze à vingt syllabe pour le corpus LSCP. En outre cette même expérience ou la reconnaissance des attitudes prosodiques implique d'avoir à traiter la fréquence fondamentale, une dimension qui n'est pas discrète, mais continue dans le temps et l'espace.

Pour traiter des structures temporelles, le réseau TRN s'appuie sur trois propriétés, qui le distingue des réseaux récurrents :

1. L'apprentissage ne s'effectue pas à partir d'un algorithme difficilement reproductible par le cerveau (rétropagation du gradient au cours du temps, Pearlmutter, 1995) ;
2. Les unités du réseau sont des intégrateurs à fuite, ce qui permet de définir la sensibilité temporelle de ce réseau ;
3. Les poids des connexions récurrentes restent fixes.

Dans le modèle original (Dominey et Ramus, 2000), l'apprentissage s'effectuait à l'aide d'une mémoire associative de la couche State vers la couche de sortie. Dans notre travail, l'apprentissage est effectué à l'extérieur du modèle pour réduire le coût informatique des

évaluations du réseau. Ces principes permettent de respecter la contrainte temporelle, par laquelle le signal analogique est transmis aux entrées du réseau, sans faire appel à une représentation symbolique.

IV.1.3.Un système unique pour le traitement des informations continues

Une partie de notre travail consiste à établir une architecture unique pour répondre à plusieurs tâches de traitement de la prosodie. Ce mécanisme a su évoluer depuis un système présent chez les primates pour traiter leurs vocalisations, qui contiennent de nombreuses variations temporelles rapides (Wang, 2000). Dans notre cas, le modèle TRN avait d'abord été conçu pour reproduire les résultats d'une tâche d'apprentissage de séquences sensori-motrices effectuée chez le primate non-humain (Dominey et coll., 1995).

Les chaînes de Markov Cachés constituent une réponse unique à de nombreux problèmes de traitement de la parole. Mais, ces modèles sont peu adaptés à l'intégration temporelle de différents types d'information sensorielle présents dans la perception et la compréhension de la parole. Ils ne proposent pas de solutions pour combiner des indices acoustiques définis sur des domaines temporels brefs (phonèmes) ou relativement longs (contour intonatifs). De plus, ces modèles nécessitent une segmentation en unité (syllabe, mot) adaptée au problème à traiter. Les simulations d'acquisition du langage basées sur des modèles connexionnistes requièrent aussi ce même type de segmentation (Christiansen et Dale, 2001). Pour la définition de modèle d'intonation au moyen de réseau récurrent (Buhmann et coll., 2000), une segmentation syllabiques est utilisée pour transmettre les informations de F0 au réseau.

L'affirmation selon laquelle des traits purement discrets constitueraient la base de la perception de catégories de phonèmes fait l'objet de sérieuses critiques (Massaro, 1987). Ladefoged (1975) concluait de ces expériences : ***« Ainsi la segmentation phonémique n'est pas la base des aptitudes linguistiques, mais leur conséquence. De toute évidence, la perception de la parole utilise une capacité de perception holistique des patterns acoustiques. En outre, nous partageons ce mode perception des séquences acoustiques avec d'autres animaux. »*** (Warren, 1994).

Notre travail nous a permis de modifier le réseau récurrent proposé par Dominey et Ramus (2000) pour qu'il puisse traiter non plus des informations discrètes, mais des dimensions continues, telles que le contour intonatif ou le spectre du signal de parole. Ainsi, ce travail ne fait plus appel à une segmentation particulière du signal de parole, mais emploie une représentation temporelle, dont l'échantillonnage est bien inférieure aux unités étudiées (phrases et mots). Ce traitement est réaliste, puisque nous tenons compte de l'écoulement naturelle du temps. Ces deux points sont désignés sous l'expression de contrainte temporelle. C'est pourquoi, nous avons recours à des intégrateurs à fuite et des connexions récurrents pour procurer au réseau une mémoire locale et globale des informations.

La remarque précédente montre une différence importante par rapport aux modèles connexionnistes récurrents, comme le réseau proposé par Elman (1990). Dans le cadre de la modélisation de l'acquisition du langage, une représentation symbolique des

données à traitées est employée (Christiansen et Dale, 2001).

IV.2.Modalité auditive

Le modèle TRN est présenté comme un modèle de traitement de la prosodie, et donc un système d'audition. Effectivement, le modèle était initialement conçu pour traiter des séquences, quelle que soit leur modalité d'origine. Nous montrons dans cette thèse que le modèle TRN fondé sur le codage par fréquence de décharges peut traiter des séquences issues de la prosodie, mais aussi des informations non verbales comme la succession de tons purs (chapitre 6 section IV.3 et IV.4).

Quelques modèles informatiques ont été proposés pour le traitement de la parole (McClelland et Elman, 1986 ; Massaro, 1987). Dans le modèle de reconnaissance de la parole de Klatt (1980), le seul niveau de la représentation auditive est le spectrogramme neural codé dans le nerf auditif.

Un certain nombre d'expériences précisent les contraintes d'un modèle de traitement auditif général. Au cours de notre travail, nous avons passé sous silence une particularité du système auditif humain : la dissociation du rythme et de l'intonation, alors que notre modèle traite les deux simultanément. Cependant, Jones et collaborateurs (Jones et coll., 1987) ont montré a plusieurs reprises que mélodie et rythme ne sont pas indépendants mais traités comme une seule dimension en perception et en mémoire. Mais, les données neurologiques ne sont pas compatibles avec cette proposition. Lorsque ces deux dimensions (rythme et mélodie) sont manipulées indépendamment l'un de l'autre, leur intégration ne se réalise pas chez le patient cérébro-lésé, ou alors de façon partielle (Peretz et coll., 2000).

Trois caractéristiques mélodiques concourent à la reconnaissance d'un air connu : le contour, les intervalles et la tonalité. Peretz et coll. (2000) ont confirmé plusieurs fois l'hypothèse que les contours mélodiques relèvent de l'hémisphère droit, et l'abstraction des intervalles de l'hémisphère gauche. La plupart des études montre bien deux mécanismes, mais chacun est capable de traiter les deux informations avec une préférence pour sa spécialité. En conséquence, une modélisation efficace du traitement de la prosodie reposerait éventuellement sur deux outils distincts capables de traiter le rythme et l'intonation en même temps, mais étant plus spécialisés l'un pour le rythme, l'autre pour l'intonation.

V.Perspectives

Trois grandes directions peuvent continuer ces travaux : la segmentation en mot du signal de parole, l'obtention d'un modèle plus complet du traitement auditif et l'application de nos méthodes à la musique.

V.1.La segmentation

Nous avons décrit dans chaque chapitre expérimental les perspectives potentielles pour chaque tâche prosodique. Nous pensons que l'ensemble de ce travail pourrait être poursuivi pour modéliser un système de traitement auditif général.

Des différences prosodiques locales permettent de définir les phrases ou les propositions (pauses, allongements des voyelles). Effectivement deux indices prosodiques, l'allongement et le contour de F0, contribuent aux processus de segmentation de la parole, et du stockage des mots d'un mini-langage artificiel (Bagou et coll., 2002). Le TRN pourrait être testé dans cette optique de segmentation du signal, pour les phrases, ou les mots. Il semble que les propriétés du TRN pour le signal de parole ou la prosodie pourraient permettre cette segmentation.

Le modèle TRN a déjà montré ces capacités de segmentation à travers la tâche de Saffran et coll. (1996). Les modifications apportées au TRN dans cette thèse suggèrent qu'une représentation spectrographique du signal de parole pourrait être utilisée pour simuler cette même expérience. En outre, cette simulation était effectuée à partir d'une représentation syllabique du flot de parole (Dominey et Ramus, 2000).

La localisation des marques du rythme semble faire appel à une procédure de segmentation (Trehub et Taylor, 1994). Cette expérience pourrait être modélisée avec le TRN, de façon à vérifier que le TRN peut percevoir des différences spectrales, de F0 et d'amplitude.

En outre, cette segmentation pourrait être employée dans le contexte de l'IAL. Cette procédure permettrait de diminuer le nombre de motifs à conserver pour caractériser une séquence de parole, lors de l'utilisation de la méthode Accumulation.

La modélisation de cette tâche de segmentation comporte d'autres intérêts. Avec un modèle informatique, il serait plus rapide de tester cette même hypothèse de segmentation sur plusieurs langues. Johnson et Jusczyk (2001) soulignent justement le manque de preuve cross-linguistique pour cette tâche. En outre, cette expérience a été répliquée avec des notes de musiques remplaçant les syllabes chez les nouveau-nés (Saffran, Johnson, Aslin et Newport, 2001). Se faisant, notre modèle pourrait être testé pour la segmentation de la musique et de la parole.

De surcroît, une simulation permettrait de tester le lien entre la segmentation et la structure prosodique des langues. Effectivement, la prosodie peut permettre de choisir le système de segmentation. Cutler (1990 et 1996) a proposé que la segmentation en mot se fonde sur des propriétés prosodiques comme la position de l'accent sur les premières syllabes des mots, pour les langues accentuelles. Les langues syllabiques ont généralement l'accent sur la dernière syllabe. Dans ce contexte, les sujets adultes confondent plus facilement des langues ayant les mêmes propriétés prosodiques pour découper le signal en mots (Bond et coll., 1998).

V.2.L'audition

La prosodie est mal définie du point de vue du traitement du signal, dans la mesure où les composantes de la prosodie (intonation, rythme, etc.) ne sont pas entièrement définies par des composantes physiques simples, telles que la fréquence fondamentale ou l'énergie. Les études psychoacoustiques (Grimault, 2000) indiquent que l'extraction de la fréquence repose sur plusieurs mécanismes différents suivant les sons à analyser. En outre, l'intonation est essentiellement basée sur les valeurs de la F0, mais l'impression qu'elle nous en laisse diverge de la réalité physique. L'étude des mécanismes d'obtention de l'intonation chez l'être humain devrait permettre d'éclaircir ces points. D'autres dimensions prosodiques, comme la qualité de la voix sont également des notions prosodiques mal définies du point de vue physique.

Pour obtenir la fréquence fondamentale, nous nous sommes appuyés sur un spectre à bande large i.e. avec une fenêtre d'analyse longue. Il est probable que des évènements brefs peuvent être détectés, mais avec une grande imprécision temporelle. Nous pourrions reconduire l'expérience de Tallal et Piercy (1973a) ainsi que d'autres expériences de psychoacoustique à partir de cette même représentation, afin de préciser la taille de cette fenêtre.

Cependant, la meilleure solution pour expliquer ce paradoxe de résolution–intégration de la fréquence fondamentale sera de travailler avec des modèles neurocomputationnels plus complets comportant des neurones impulsionnels et des synapses dynamiques (Denham, 1999).

Comme pour le rythme face aux différents tempos, un processus d'abstraction est contenu dans la perception de la hauteur. Effectivement les nourrissons semblent traiter la hauteur suivant un contour global (Trehub et Trainor, 1994). Le mécanisme d'abstraction proposé par Dominey et Ramus (2000) pour répliquer l'expérience de Marcus et coll. (1999) pourrait effectuer un traitement de la hauteur en respectant cette particularité. Le contour apparaît alors sous la forme d'une règle abstraite de formation de la mélodie, à partir d'une hauteur absolue.

Certaines modifications pourraient être apportées comme une adaptation des récepteurs auditifs aux stimuli présentés (Mercado III et coll., 2001). En outre, nous ne tenons pas compte d'une dissociation du traitement de la prosodie et des informations spectrales.

V.3.La musique

Les expériences ayant trait à l'IAL ont montré que le réseau TRN pouvait distinguer les langues en fonction de leurs classes rythmiques. Cette méthode ne permet pas de retrouver les phonèmes, mais donne une caractérisation globale des trajectoires de la F0 et des formants.

Le rythme est mieux défini pour la musique, que pour la parole. Pour extraire le rythme d'une pièce musicale, il faut être en mesure d'extraire des indices rythmiques, comme un brusque changement d'intensité. Un certain nombre de règles ont été avancées pour expliquer la détection des événements sonores dans un flot continu (Bregman, 1994, *cf.* chapitre 2 section II.1.1). L'architecture du réseau TRN doit permettre

de respecter une partie de ces règles. Une expérience simple, menée avec le TRN doit pouvoir montrer que des sons démarrant à des instants différents peuvent être isolés. En outre, la condition de programmation de la transformation (effet rendu par le vol du Bourdon) doit pouvoir être reproduite (expérience de Van Noorden, 1975), sachant que deux sons voisins en fréquence donnent des vecteurs proches après l'encodage du TRN. Ceci est rendu par la nature gaussienne du codage de F0, et pourrait être accentué en modifiant la distribution des connexions entre la couche d'entrée, et les couches internes.

Des expériences comportementales permettent de poser certaines contraintes sur le mécanisme de perception. Par exemple, les sujets semblent privilégier une représentation par contour de la mélodie (Deutsch, 1980). Cette expérience pourrait être simulée avec le TRN, pour vérifier notre mécanisme.

L'étape suivante pour percevoir le rythme est de représenter une séquence des marques rythmiques. Les auditeurs humains perçoivent et utilisent les hiérarchies pour mémoriser des séquences musicales. Il conviendrait de tester le modèle pour des types simples et complexes de mélodies (Deutsch, 1980). Lorsque la surface musicale coïncide avec la structure hiérarchique de la mélodie, les performances du modèle doivent être optimum. Cependant, il est difficile de dire si ce phénomène peut être retrouvé sans modifier le modèle TRN.

Pour finaliser un modèle de traitement du rythme chez l'être humain, il devra être insensible aux variations de tempo. Cette particularité semble être celle d'un modèle assez simple, dans la mesure où même les oiseaux sont insensibles à ces variations. Une première expérience, effectuée avec les attitudes prosodiques, montre que le réseau est peu sensible au ralentissement « artificiel » de la parole, mais des expériences beaucoup plus complètes devraient être envisagées pour étudier cette propriété.

La musique tient également une place importante dans le « parler bébé » en terme de mélodie : les tierces, les quartes, les quintes et les octaves (Fernald, 1976) y sont prédominantes. Ce rapport original entre musique et parole est probablement lié au système auditif, et plus particulièrement au mécanisme de perception de la hauteur. Des contraintes biologiques de traitement peuvent expliquer le choix de certains rapports fréquentiels (octave et quinte), qui apparaissent plaisant à l'oreille (Cariani, 1999). Les rapports entre musique et acquisition du langage (Dodane, 2003 ; Saffran et coll., 1999) pourraient faire l'objet d'études approfondies, en tenant compte d'un modèle unique pour simuler des tâches de traitement avec de la parole et de la musique.

VI.Conclusion

Notre travail s'organise autour de structures temporelles définies suivant différentes échelles inscrites dans un Continuum Temporel. Nous avons étudié un unique mécanisme dédié au traitement de séquences, pour vérifier qu'il pouvait traité plusieurs des échelles représentatives de l'Hypothèse de Continuum Temporel. Effectivement, le réseau récurrent temporel TRN a pu executer quatre tâches de traitement de la prosodie :

1. L'Identification Automatique des Langues (IAL) indique la langue parlée à partir du signal acoustique , qui forme une structure temporelle globale ;
2. L'Identification des Attitudes Prosodiques permet de tester les facultés du réseau pour le traitement du contour intonatif de phrases courtes ;
3. La distinction entre ces catégories lexicales s'opère sur les mots et interviendrait dans un processus d'acquisition de la syntaxe. Nos études ont prouvé que des indices temporels, contenus dans le contour intonatif contribuaient à la discrimination entre mots de fonction et mots de contenu, pour deux langues.
4. De surcroît, en perturbant le réseau TRN, nous montrons qu'il ne peut plus effectuer correctement cette tâche de catégorisation lexicale, et qu'il exhibe un profil semblable à celui des enfants SLI lors du traitement des évènements auditifs rapides, ce qui procure une simulation du comportement d'au moins un sous-groupe des enfants SLI, qui présentent des troubles de la syntaxe.

En outre, ce mécanisme respecte une contrainte temporelle, puisque le réseau emploie

un échantillonnage très inférieur aux unités des données traitées. Il ainsi peut traiter directement une représentation spectrographique de la prosodie du signal de parole pour discriminer des langues de classes rythmiques différentes, et distinguer les mots de fonction des mots de contenu. Ainsi, le signal acoustique n'est pas représenté par des séquences d'éléments discrets ou symboles, dont la durée est spécifiée en fonction de l'échelle à traiter, mais est considéré sous sa forme analogique.

Bibliographie

Abe, T., Kobayashi, T. et Imai, S. (1995). *Harmonics tracking and pitch extraction based on instantaneous frequency*. Proc. IEEE-ICASSP, pp. 756-759.

Abney, S. P. (1987). *The english noun phrase and its sentential aspect*. Thèse, MIT.

Ahissar, E., Haidarliu, S. et Zacksenhouse, M. (1997). *Decoding temporally encoded sensory input by cortical oscillations and thalamic phase comparators*. Proc Natl Acad Sci, pp. 11633-11638, USA.

Alexandre, F., Guyot, F., Haton, J. P. et Burnod, Y. (1991). The cortical column : a new processing unit for multilayered networks. *Neural Networks,* 4.

Allen, J. F. (1991). Time and Time Again: The Many Ways to Represent Time The International. *Journal of Intelligent Systems,* 6(4): 341-355.

Almeida, L. (1987). *A learning rule for asynchronous perceptrons with feed- back in a combinatorial environment.* Proceedings of the IEEE First international Conference on Neural Networks, pp. 609-618.

André-Obrecht, R. (1988). *A New Statistical Approach for Automatic Speech Segmentation*. IEEE Trans. on ASSP.

Andruski, J. E. et Kuhl, P. K. (1996). *The acoustic structure of vowels in mother's speech to infants and adults*. Proceedings. Fourth International Conference on Spoken Language Processing.

Arvaniti, A., Ladd, D. et Mennen, L. (1998). Stability of tonal alignment: the case of

Greek prenuclear accents. *Journal of Phonetics,* 26: 3-25.

Aslin, R. N. et Smith, L. B. (1988). Perceptual development. *Annual review of Psychology,* 39: 435-473.

Aubergé, V., Grépillat, T. et Rilliard, A. (1997). *Can we perceive attitudes before the end of sentences ? The gating paradigm for prosodics contours.* Eurospeech, Proceedings of the European Conference on Speech Communication and Technology, pp. 871-874, Rhodes, Greece.

Back, A. D. et Tsoi, A. C. (1993). A simplified gradient algorithm for iir synapse multilayer perceptrons. *Neural Computation,* 5: 456-462.

Bagou, O., Fougeron, C. et Frauenfelder, U. H. (2002). *Contribution of Prosody to the Segmentation and Storage of " Words " in the Acquisition of a New Mini-Language*. Speech Prosody 2002, pp. 15-162.

Bahrick, L. E. et Pickens, J. N. (1988). Classification of bimodal English and Spanish language passages by infants. *Infant Behavior and Development,* 11: 277-296.

Baker, C. et Rao, R. (Soumis). Analysis of a dynamic synapse model for supervised speech learning.

Barnard, E., Cole, R. A., Vea, M. P. et Alleva, F. A. (1991). *Pitch detection with a neural-net classifier*. IEEE Trans. Signal Process., pp. 298-307.

Barone, P. et Joseph, J. P. (1989). Prefrontal cortex and spatial sequencing in the macaque monkey. *Experimental Brain Research,* 78: 447-464.

Barto, A. G. (1995). Adaptive critic and the basal ganglia. *Models of information processing in the basal ganglia*. J. C. Houk, J. L. Davis et D. G. Beiser (Eds.), Cambridge, MIT Press., pp.215-232.

Bates, E. et MacWhinney, B. (1989). Functionalism and the competition model. *The Crosslinguistic Study of Language Processing*. B. MacWhinney et E. Bates (Eds.), New York, Cambridge University Press.

Beiser, D. (1995). *Models of Information Processing in the Basal Ganglia*. D. Beiser (Ed.), Cambridge, MA: MIT Press.

Beiser, D. G. et Houk, J. C. (1998). Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology,* 79: 3168-3188.

Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A., Masure, M.C., Samson, Y (1998). Lateralization of speech and auditory temporal processing. *Journal of Cognitive Neuroscience,* 10: 536-540.

Benasich, A. A. et Tallal, P. (1996). Auditory temporal processing thresholds, habituation, and recognition memory over the 1st year. *Infant Behavior and Development,* 19 (3): 339-357.

Benasich, A. A. et Tallal, P. (2002). Infant discrimination of rapid auditory cues predicts later language impairment. *Behavioural Brain Research,* 136: 31-49.

Bengio, Y., Simard, P. et Frasconi, P. (1994). *Learning long-term dependencies with gradient descent is difficult.* IEEE Transactions on Neural Networks, pp. 157-166.

Berkling, K. M. (1996). *Automatic Language Identification with Sequences of*

Language-Independent Phoneme Clusters. Thèse, Oregon Graduate Institute of Science & Technology.

Berns, G. S. et Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience,* 10: 108-121.

Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language,* 11: 557-578.

Bernstein-Ratner, N. (1986). Durational cues which mark clause boundaries in mother-child speech. *Journal of Phonetics,* 14(2): 303-309.

Béroule, D. (1985). *Un modèle de mémoire adaptative, dynamique et associative pour le traitement automatique de la parole*. Thèse, Universitéde Paris XI, Orsay.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., et Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General,* 117(1): 21-33.

Besson, M., Magne, C. et Schön, D. (2001). *Apport de la méthode des potentiels évoqués à l'étude du traitement prosodique*. Journées Prosodie, pp. 19-34, Grenoble.

Bijeljac-Babic, R., Bertoncini, J. et Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances? *Developmental Psychology,* 29: 711-721.

Bishop, D. (1979). Comprehension in developmental language disoders. *Developmental Medicine and Child Neurology,* 21: 225-238.

Bishop, D. V. M. (1997). *Uncommon understanding: development and disorders of language comprehension in children*. Hove: Psychology Press.

Bishop, D. V. (1998). Development of the Children's Communication Checklist (CCC): a method for assessing qualitative aspects of communicative impairment in children. *J Child Psychol Psychiatry,* 39(6): 879-91.

Bishop, D. V. M. (2000). How does the brain learn language? Insights from the study of children with and without language impairment. *Developmental Medicine and Child Neurology,* 42: 133-142.

Bishop, D. V. M., Carlyon, R. P., Deeks, J. M. et Bishop, S. J. (1999). Auditory temporal processing impairment: neither necessary nor sufficient for causing language impairment in children. *Journal of Speech, Language and Hearing Research,* 42: 1295-1310.

Bishop, D. V. M., Bishop, S. J., Bright, P., James, C., Delaney, T., Miller, S. et Tallal, P. (in press). Different origin of auditory deficit and weak phonological short-term memory in children with specific language impairment: evidence from a twin study. *Journal of Speech, Language and Hearing Research.*

Blackburn, C. C. et Sachs, M. B. (1990). The representation of the steady-state vowel /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology,* 63: 1191-1212.

Blackwell, A. W. et Bates, E. (1995). Inducing agrammatic profiles in normals: Evidence for the selective vulnerability of morphology under cognitive resource limitation. *Journal of Cognitive Neuroscience,* 7: 228-257.

Blanc, M. (1981). *Etude nosographique de 6990 handicaps, maladies chroniques et/ou malformations chez 5192 enfants présentant 4037 atteintes isolées et 2953 atteintes*

associés. Thèse, Thèse de médecine de l'université Claude-Bernard Lyon I.

Blanc, J.-M., Dominey, P.F. (2001). *La prosodie comme lien entre une tâche de discrimination lexicale et un trouble du langage: simulation par un réseau récurrent temporel*. Actes des Journées Prosodie, Grenoble 2001, pp. 161-164.

Blanc, J.-M., Dominey, P.F. (2002). *Temporal processing and prosodic bootstrapping of syntax: A simulation study*. Proceedings of the ISCA Workshop on Temporal Integration in the Perception of Speech, pp. 61, Aix-en-Provence.

Blanc, J.-M., Dodane, C., Dominey, P.F. (2003b). *Prosodic cues for lexical categorization : Simulation and Data*. 2003 Convention of American Speech-Language Hearing (CD-ROM).

Blanc, J.-M. et Dominey, P. F. (2003). Identification of prosodic attitudes by a temporal recurrent network. *Cognitive Brain Research,* 17(3): 693-699.

Blanc, J.-M. et Dominey, P. (2004). *On using prosodic cues for the identification fo content and function words*. Speech Prosody 2004, Nara, Japon.

Blanc, J.-M., Dodane, C. et Dominey, P. F. (2003a). *Temporal Processing for Syntax Acquisition: A simulation study*. the 25th Annual Conference of the Cognitive Science Society, pp. 145-150.

Blanc, J.-M., Dodane, C. et Dominey, P. F. (En préparation). Neural network processing of Natural language II : Prosodic Boostrapping of lexical Categories. *Language and Cognitive Process*.

Boersma, P. (1993). *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. Proceedings of the Institute of Phonetic Sciences, pp. 97-110.

Bond, Z. S. et Fokes, J. (1991). *Identifying foreign languages*. In proceedings of the XIIth international congress of phonetic sciences, pp. 198-201.

Bond, Z. S., Stockmal, V. et Muljany, D. (1998). Learning to identify a foreign language. *Language Science,* 20(4): 353-367.

Bosch, L. et Sebastián-Gallés, N. (1997). Native language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition,* 65: 33-69.

Bradlow, A. R., Kraus, N., Nicol, T. G., McGee, T. J., Cunningham, J., S.G., Z. et Carrell, T. D. (1999). Effects of lengthened formant transition duration on discrimination and neural representation of synthetic CV syllables by normal and learning-disabled children. *Journal of the Acoustical Society of America,* 106(4): 2086-2096.

Braine, M. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development,* 41 (Serial n° 164).

Braitenberg, V. (1967). Is the cerebellar cortex a biological clock in the millisecond range? *Progress in Brain Research*. **25**, pp.334-336.

Breazeal, C. (2000). *Sociable machines :expresive social exchange between human and robots*. Thèse, MIT lab.

Bregman, A. S. (1994). L'analyse des scènes auditives : l'audition dans des environnements complexes. *Penser les sons : Psychologie cognitive de l'audition*. S.

McAdams et E. Bigand (Eds.), Paris, PUF, pp.11-40.

Brown, J., Bullock, D. et Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience,* 19(23): 10502-11.

Buhmann, J., Vereecken, H., Fackrell, J., Martens, J. et Coile, B. (2000). *Data driven intonation modelling of 6 languages*. Proceedings of ICSLP, pp. 179-182, Beijing, China.

Buonomano, D. V. (2000). Decoding Temporal Information: A Model Based on Short-Term Synaptic Plasticity. *Journal of Neuroscience,* 2000.

Buonomano, D. V. et Mauk, M. D. (1994). Neural network model of the cerebellum: Temporal discrimination and the timing of motor responses. *Neural Computation,* 6: 38-55.

Buonomano, D., V. et Merzenich, M., M (1995). Temporal Information Transformed into a Spatial Code by a Neural Network with Realistic Properties. *Science,* 267: 1028-1030.

Buonomano, D. V. et Merzenich, M. M. (1999). A neural network model of temporal code generation of position invariant pattern recognition. *Neural Computation,* 11: 103-116.

Buonomano, D. V. et Karmarkar, U. R. (2002). How do we tell time? *Neuroscientist,* 8: 42-51.

Burnham, D. et Torstenson, C. (1995). *The development of phonological bias: perception and production of swedish vowels and tons by english speakers*. Proceedings of XIIIth international congress of phonetic sciences, pp. 558-561.

Burnod, Y. (1988). *An Adaptative Neural Network: The Cerebral Cortex*. Paris, Masson.

Burns, E. M. et Ward, W. D. (1982). Intervals, scales and tunings. *Psychology of music*. D. Deutsch (Ed.), Orlando, Florida, Academic Presss, pp.241-269.

Campione, E. (2001). *Quelques outils pour l'étiquetage prosodique des corpus oraux*. Journées Prosodie, pp. 103-106, Grenoble.

Campione, E. et Véronis, J. (1998). *A multilingual prosodic database*. Proc. of ICSLP'98, Sidney.

Caradan, L. (2001). *la prosodie professionnelle de l'enseignant en primaire*. Journées Prosodie, pp. 149-152, Grenoble.

Cariani, P. (1999). Neural timing nets for auditory computation. *Computational Models of Auditory Function*. S. Greenberg et M. Slaney (Eds.), IOS Press.

Cariani, P., Tramo, M. et Delgutte, B. (1997). *Neural representation of pitch through temporal autocorrelation*. Proceedings, Audio Engineering Society Meeting (AES), New York.

Caseiro, D. et Trancoso, I. M. (1998). *Identification of Spoken European Languages*. Proceedings European Signal Processing Conference (Eusipco-98), Rhodes, Greece.

Cassidy, K. W. et Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language,* 30: 348-369.

Chomsky, N. (1965). *Aspect of the theory of syntax*. Cambridge , MA, MIT Press.

Chomsky, N. (1975). *Reflections on Language*. Pantheon.

Christiansen, M. H. et Dale, R. A. C. (2001). *Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition*. Proceedings of the 23rd Annual Conference of the Cognitive Science Society, pp. 220-225, Mahwah, NJ, Lawrence Erlbaum.

Clark, M. G., Rosen, G. D., Tallal, P. et Fitch, R. H. (2000). Impaired processing of complex auditory stimuli in rats with induced cerebrocortical microgyria: An animal model of developmental language disabilities. *Journal of Cognitive Neuroscience,* 12(5): 828-39.

Cleeremans, A. (1993). *Mechanics of Implicits Learnig Connectionist Models of Sequence Processing*. Cambridge, Massachusetts, The MIT Press.

Cleeremans, A. et McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General,,* 120(3): 235-253.

Combrinck, H. P. (1999). *A cost, complexity and performance comparison of two automatic language identification architectures*. Thèse, University of Pretoria.

Cowie, R., Douglas-Cowie, E., Tsapatso lis, N , Votsis, G , Kollias, S , Fellenz, W. et Taylor, J. G. (2001). *Emotion recognition in human-computer interaction*. IEEE Sig Proc Mag 18(1), 32-80.

Creutzfeldt, O., Ojemann, G. et Lettich, E. (1989). Neuronal activity in the human lateral temporal lobe. I. Responses to speech. *Experimental Brain Research,* 77: 451-475.

Cummins, F., Gers, F. et Schmidhuber, J. (1999). Automatic discrimination among languages based on prosody alone. *Instituto Dalle Molle di studie sull'Intelligenza Artificiale, Switzerland Février. 99.*

Cummins, F., Gers, F. et Schmidhuber, J. (1999). Comparing Prosody accross many languages. *Instituto Dalle Molle di studie sull'Intelligenza Artificiale, Switzerland Juillet. 99.IDSIA Technical report Juillet 1999*.

Cummins, F., Gers, F. et Schmidhuber, J. (1999). *Language identification from prosody without explicit features*. in: Proc. EUROSPEECH 99, pp. 371-374.

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. G. Altmann (Ed.), Cambridge Mass, MIT Press, pp.105-121.

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research,* 22: 109-131.

Cutler, A. (1996). Prosody and the word boundary problem. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. J. L. Morgan et K. Demuth (Eds.), Mahwah, NJ, Lawrence Erlbaum Associates, pp.87-99.

Cutler, A. et Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language,* 2: 133-142.

Cutler, A. et Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance,* 14: 113-121.

D'Amato (1988). A search for tonal pattern perception in cebus monkeys: why monkeys can't hum a tune. *Music Perception,* 5: 453-480.

Dayan, P. et Sejnowski, T. (1994). TD(lambda) Converges with Probability 1. *Machine Learning, 3, 9-44,* 14: 295-301.

De Cheveigné, A. (1991). *Speech f0 extraction based on Licklider's pitch perception model*. Proc. ICPhS, pp. 218-221.

de Pijper, J. R. (1983). *Modelling British English intonation*. Dordrecht - Holland, Foris.

Dean, T. et Kanazawa, K. (1988). *Probabilistic Temporal Reasoning*. Proc. of the National Conference of the American Association for Artificial Intelligence (AAAI88), Morgan Kaufman.

Dehaene, S. et Changeux, J. P. (1989). A simple model of prefontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience,* 1.

Dehaene-Lambertz, G. (1995). *Capacités linguistiques précoces et leurs bases cérébrales*. Thèse, Université Paris VI, Paris.

Dellaert, F., Polzin, T et Waibel, A. (1996). *Recognizing Emotion in Speech*. Fourth International Conference on Spoken Language Processing 3, pp. 1970-1973.

Demany, L. et Armand, F. (1984). The perceptual reality of tone chroma in early infancy. *Journal of the Acoustical Society of America,* 76: 57-66.

Demany, L., McKenzie, B. et Vurpillot, E. (1977). Rhythm perception in early infancy. *Journal of Acoustic Society of America,* 76: 57-66.

Demuth, K. (1996). The prosodic structure of early words. *Signal to syntax: Boostrapping from speech to grammar in early acquisition*. J. L. Morgan et K.Demuth. (Eds.), pp.171-186.

den Os, E. A. (1988). *Rhythm and tempo of Dutch and Italian, a constrative study*. Thèse, Ultrecht University.

Denham, S. L. (1999). Cortical Synaptic Depression and Auditory Perception. *Computational Models of Auditory Function*. S. Greenberg et M. Slaney (Eds.), Amsterdam, NATO ASI Series, IOS Press.

Deutsch, F. (1980). The processing of structured and unstructured tonal sequences. *Perception and psychophysics,* 28: 381-389.

Dobie, R. et Berlin, C. (1979). Influence of otis media on hearing and development. *Annals of Otology, Rhinology and Laryngology,* 88 (supplement 60): 48-53.

Dodane, C. (2003). *La langue en harmonie : Influences de la formation musicale sur l'apprentissage précoce d'une langue étrangère*. Thèse, université de franche-comté.

Dodane, C., Blanc, J.-M. et Dominey, P. F. (En préparation). Prosodic structure of Open and Closed Class Words in CDS in English, French and Japanese. *Journal of Child Language*.

Doerksen, K. (2000). Pulsed Neural Networks: Temporal Signal Processing Using Artificial Neural Networks with Dynamic Synapses. *CYSF 2000 Report.*

Dominey, P. F. (1995). Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics,* 73: 265-274.

Dominey, P. F. et Arbib, M. A. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex,* 2: 153-175.

Dominey, P. F. et Ramus, F. (2000). Neural network processing of natural language: I.

Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes,* 15(1): 87-127.

Dominey, P. F., Arbib, M. A. et Joseph, J. P. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *Journal of Cognitive Neuroscience,* 7: 311-336.

Dominey, P. F., Hoen, M., Blanc, J.-M. et Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies. *Brain and Language,* 86: 207-225.

Doval, B. (1994). *Estimation de la fréquence fondamentale des signaux sonores*. Thèse, LAFORIA, Universite Paris 6, Paris, France.

Drake, C. (2002). *Temporal organisation of sound sequences: Same processes in music, environmental scenes and speech ?* ISCA Workshop "Temporal Integration in the Perception of Speech", Aix en Provence.

du Preez, J. A. et Weber, D. M. (1998). *Efficient highorder hidden Markov modelling*. Proceedings 5th International Conference on Spoken Language Processing, pp. 2911--2914, Sydney, Australia.

Duifhuis, H., Willems, L. F. et Sluyter, R. J. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *Journal of the Acoustical Society of America,* 71: 1568-1580.

Durand, S. (1995). *TOM, une architecture connexionniste de traitement de séquences. Application à la reconnaissance de la parole*. Thèse, Univ. Henri-Poincaré, Nancy I.

Durieux, G. et Gillis, S. (2000). Predicting grammatical classes from phonological cues: An empirical test. *Approaches to bootstrapping: phonological, syntactic and neurophysiological aspects of early language acquisition*. B. Höhle et J. Weissenborn (Eds.), Amsterdam, Benjamins, pp.189-232.

Dutat, M. (2000). *Caractérisation de la langue parlée par modèles de séquences d'événements acoustiques*.

Elis Weismer, S. et Hesketh, L. (1993). The influence of prosodic and gestural cues on novel word acquisition by children with specific language impairment. *Journal of Speech and Hearing Research,* 36: 1013-1025.

Elis Weismer, S. et Hesketh, L. (1996). Lexical learning by children with specific language impairment : Effects of linguistic input presented at varying speaing rates. *Journal of Speech and Hearing Research,* 39: 177-190.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science,* 14: 179-211.

Eyer, J. et Leonard, L. (1995). Functional categories and specific language impairment: A case study. *Language Acquisition,* 4: 177-203.

Fahlman, S. E. (1991). The recurrent cascade-correlation architecture. *Technical Report CMU-CS-91-100, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213 / San Mateo, CA.*

Fant, G., Kruckenberg, A. et Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. *Journal of phonetics,* 19: 351-365.

Farinas, J. (2002). *Une modélisation automatique du rythme pour l'identification des langues*. Thèse, Université Toulouse III Paul Sabatier (Spécialité Informatique) ITI

UMR 5505.

Farinas, J. et André-Obrecht, R. (2000). *Identification Automatique des langues : variation sur les multigrammes,*. Actes XXIIIèmes Journées d'Etude sur la Parole, JEP'2000, Aussois, 19-23 juin 2000.

Fellbaum, C., Miller, S., Curtiss, S. et Tallal, P. (1995). *An auditory processing deficit as a possible source of SLI.* Proceedings of the 19th Annual Boston University conference on Language Development, pp. 204-215, Boston.

Fernald, A. (1976). The mother's speech to the newborn. *Paper presented at the Max-Plank Institute for Psychiatry, Munich (non published) (Dodane, 2003)*.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development,* 8: 181-195.

Fernald, A. (1989). Intonation and communication intent in mother's speech to infants: is the melody the message ? *Child Development,* 60: 1497-1510.

Fernald, A. et Kuhl, P. (1981). *Fundamental frequency as an acoustic determinant of infant preference for " Motherese "*. Paper presented at the Biennal Meeting of the Society for Research in Child Development., Boston.

Fernald, A. et Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology,* 20(1): 104-113.

Fernald, A. et Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development,* 10: 279-293.

Fernald, A. et McRoberts, G. (1996). Prosodic bootstrapping: A ciritical analysis of the argument and evidence. *Signal to syntax: Boostrapping from speech to grammar in early acquisition*. J. L. Morgan et K. Demuth (Eds.), Mahwah, NJ, Lawrence erlbaum associates, pp.365-388.

Fiez, J. A., Raichle, M. E., Miezin, F. M., Petersen, S. S., Tallal, P. et Katz, W. F. (1995). PET studies of auditory and phonological processing: effects on stimulus characteristics and task design. *Journal of Cognitive Neuroscience,* 7, 357-375.

Fiez, J. A., Raife, E. A., Balota, D. A., Schwarz, J. P., Raichle, M. E. et Petersen, S. E. (1996). A positron emission tomography study of the short-term maintenance of verbal information. *Journal of Neuroscience,* 16: 802-822.

Finch, S. et Chater., N. (1992). *Bootstrapping syntactic categories*. In Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society of America, pp. 820-825.

Fisher, C. et Tokura, H. (1996). Prosody in speech to infants: Direct and indirect cues to syntactic structure. *Signal to syntax*. J. Morgan et K. Demuth (Eds.), Lawrence Erlbaum Assoc, pp.343-363.

Fitch, H., Brown, C., O'Connor, K., Tallal, P. (1993). Functional lateralization for auditory temporal processing for male and female rats. *Behavioral Neuroscience,* 107: 844-850.

Fletcher, P. (1983). *From sound to syntax : A learner's guide*. Proceedings from the Wisconsin Symposium on research in Child Language Disorders, pp. 1-31, Madison, University of Wisconsin.

Fraisse (1967). *Psychologie du temps*. Paris, PUF, 2 èdition.

Fraisse (1974). *La psychologie du rythme*. Paris, PUF.

Francis, W. et Kucera., H. (1982). Frequency Analysis of English Usage: Lexicon and Grammar. *Boston: Houghton-Mifflin.*

Funahashi, S., Bruce, C. J. et Goldman-Rakic, P. S. (1989). Mneumonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology,* 61: 331-349.

Funahashi, S., Bruce, C. J. et Goldman-Rakic, P. S. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *Journal of Neurophysiology,* 63: 814-831.

Funahashi, S., Inoue, M. et Kubota, K. (1993). Delay-related activity in the primate prefrontal cortex during sequential reaching tasks with delay. *Neuroscience Research,* 18: 171-175.

Fuster, J. M. (1985). The prefrontal cortex and temporal integration. *Cereb. Cortex,* 4: 151-177.

Fuster, J. M. (1993). Frontal lobes. *Current Opinion in Neurobiology,* 3: 160-165.

Fuster, J. M. et Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science,* 173: 652-654.

Galaburda, A. M. (1994). *Evidence from aberrant auditory anatomy in developmental dyslexia*. Proceedings of the national academy of sciences of the united states of America, 91, 8010-8013.

Galvès, A., Garcia, J. E., Duarte, D. et Galves, C. (2002). *Sonority as a Basis for Rhythmic Class Discrimination*. Proceedings of Speech Prosody 2002.

Gerken, L. A. (1994). Sentential processes in early child language: Evidence from the perception and production of function morphemes. *The transition from speech sounds to spoken words*. H. C. Nusbaum et J. C. Goodman (Eds.), Cambridge, MA, MIT Press, pp.271-298.

Gerken, L. A. et McIntosh, B. J. (1993). The interplay of function morphemes and prosody in early language. *Developmental Psychology,* 29: 448-457.

Gerken, L., Jusczyk, P. W. et Mandel, D. (1994). When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition,* 51(3): 237-265.

Gibbon, J., Malapani, C., Dale, C. L. et Gallistel, C. R. (1997). Toward a neurobiology of temporal cognition: advances and challenges. *Current Opinion in Neurobiology,* 7: 170-184.

Giles, C. L. et Horne, B. G. (1994). *Representation and learning in recurrent neural network architectures.* In Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems, pages 128-134, Center for Systems Science, Dunham Laboratory, Yale University New Haven, CN. Center for Systems Science, Dunham Laboratory.

Gillies, A. et Arbuthnott, G. (2000). Computational Models of the Basal Ganglia. *Movement Disorders,* Vol. 15, No. 5: 762-770.

Gleason, T. et Bharucha, J. (1990). *Speech accompanied by a tone with aligned or misaligned stress*. Paper presented at the 31st Annual Meeting of the Psychonomic Society, New Orleans, LA.

Gleitman, L. et Wanner, E. (1982). The state of the state of the art. *Language acquisition: The state of the art*. E. W. L. Gleitman (Ed.), Cambridge, UK, Cambridge University Press, pp.3-48.

Gleitman, L., Gleitman, H., Landau, B. et Wanner, E. (1988). Where learning begins. *The Cambridge Linguistic Survey*. F. Newmeyer (Ed.), New York, Cambridge University Press. **3**, pp.150-193.

Goldberg, M. E. et Colby, C. L. (1989). The neurophysiology of spatial vision. *Handb. Neuropsychol.,* 2: 301–315.

Goldinger, S. D., Pisoni, D. B. et Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology:Learning, Memory and Cognition,* 17: 152-162.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Handbook of Physiology. The Nervous System. Higher Functions of the Brain*, Bethesda, MD, Am. Physiol. Soc, sect. 1. **5**, pp.373-417.

Goldman-Rakic, P. S. (1995). Toward a circuit model of working memory and the guidance of voluntary motor action. *Models of Information Processingin the Basal Ganglia, edited by J. C. Houk, J. L. Davis, and D. G. Beiser. Cambridge: MIT Press, , p. 131-148.*

Gopnik, M. (1990a). Feature-blind grammar and dysphasia. *Nature,* 344: 715.

Gopnik, M. (1990b). Feature blindness : a case study. *Language Acquisition,* 1: 139-164.

Grabe, E. et Low, E. L. (2002). Durational Variability in Speech and the Rhythm Class Hypothesis. *Papers in Laboratory Phonology 7, Mouton.*

Grethe, J. S. et Arbib, M. A. (2001). *Computing the brain: A Guide to Neuroinformatics*. M. A. Arbib et J. S. Grethe (Eds.).

Grimault, N. (2000). *Perception de la hauteur des sons complexes harmoniques: étude des mécanismes sous-jacents et relation avec l'analyse de scènes auditives*. Thèse, Université Claude Bernard Lyon 1, Lyon.

Guilfoyle, E. et Noonan, M. (1992). Functional Categories and Language Acquisition. *Canadian Journal of Linguistics,* 37: 241-272.

Guilfoyle, E., Allen, S. et Moss, S. (1991). *Specific language impairment and the maturation of functional categories*. Boston University Conference on Language Development, Boston.

Haarman, H., Just, M. et Carpenter, P. (1997). Aphasic sentence comprehension as a resource deficit : A computationnal approach. *Brain and Language,* 59: 76-120.

Habib, M. (2002). *Phonology, phonetics, and temporal processing in developmental dyslexia: From mechanisms to remediation*. Proceedings of Temporal integration in the perception of speech, Aix-en-Provence.

Hamon, C., Moulines, E. et Charpentier, F. (1989). *A diphone system based on time-domain prosodic modifications of speech*. Proc. ICASSP 89, pp. 238-241.

Harel, D. (1979). *First-Order Dynamic Logic*. New York, Springer Verlag.

Harrington, D. L., Haaland, K. Y. et Knight, R. T. (1998a). Cortical networks underlying mechanisms of time perception. *Journal of Neuroscience,* 18: 1085-1095.

Harrington, D. L., Haaland, K. Y. et N., H. (1998b). Temporal processing in the basal ganglia. *Neuropsychology,* 12: 3-12.

Hartley, D. E. H. et Moore, D. R. (2001). *Deficits in processing efficiency in individuals with SLI and dyslexia*. Proceedings of Sensory bases of reading and language disorders.

Hauser, M. D. (2002). Qu'y a-t-il de si spécial dans la parole ? *Les langages du cerveau*. E. Dupoux (Ed.), Paris, Edition Odile Jacob.

Haynes, C. (1982). *Vocabulary acquisition problems in language disordered children*. Thèse, University of London, Guys hospital medical school.

Hazeltine, E., Grafton, S. T. et Ivry, R. (1997). Attention and stimulus characteristics determine the locus of motor sequence encoding: a PET study. *Brain,* 120: 123-140.

Hazen, T. J. et Zue, V. W. (1994). *Recent improvements in an approach to segment-based automatic language identification*. In Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP 94), Yokohama, Japan, September 1994.

Hazen, T. J. et Zue, V. W. (1997). Segment-based automatic language identification. *Journal of the Acoustical Society of America,* 101(4): 2323-2331.

He, J., Hashikawa, T., Ojima, H. et Kinouchi, Y. (1997). Temporal integration and duration tuning in the dorsal zone of the cat auditory cortex. *Journal of Neuroscience,* 17: 2615-2626.

Henderson, B. (1978). *Older language impaired children's processing of rapidly changing acoustic signals.* Convention of the American Speech-Language-Hearing Association, San Francisco.

Hérault, J. et Jutten, C. (1994). *Réseaux neuronaux et traitement du signal*. Grenoble, Hermès, INPG.

Hermes, D. J. (1993). Pitch Analysis. *Visual Representations of Speech Signals*. M. Cooke, S. Beet et M. Crawford (Eds.), New York, John Wiley & Sons, pp.3-25.

Hess, W. J. (1983). *Pitch Determination of Speech Signals*. Berlin, Springer-Verlag.

Hess, W. J. (1992). Pitch and voicing determination. *Advances in Speech Signal Processing*. S. Furui et M. M. S. M. Dekker (Eds.), New York, pp.3-48.

Hikosaka, O. (1989). Role of basal ganglia in initiation of voluntary movements. *Dynamic Interactions in Neural Networks: Models and Data*. M. A. Arbib et S. Amari (Eds.), Berlin, Springer-Verlag, pp.153-167.

Hikosaka, O. et Wurtz, R. H. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. III. Memory-contingent visual and saccade responses. *Journal of Neurophysiology,* 49: 1268-1284.

Hikosaka, O., Sakamoto, M. et Sadanari, U. (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *Journal of Neurophysiology,* 61: 814-832.

Hill, A. V. (1936). Excitation and accomodation in nerve. *Proc. R. Soc. London B.,* 119:

305-355.

Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence,* 63: 305-340.

Hirst, D. et Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix,* 15: 71-85.

Hirst, D. et Di Cristo, A. (1998). *Intonation Systems. A Survey of Twenty Languages*. D. Hirst et A. Di Cristo (Eds.).

Hochreiter, J. (1991). *Untersuchungen zu dynamischen neuronalen Netzen.* Thèse, Master's thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München.

Hochreiter, S. et Schmidhuber, J. (1997a). Long short-term memory. *Neural Computation,* 9(8): 1735-1780.

Hochreiter, S. et Schmidhuber, J. (1997b). Lstm can solve hard long time lag problems. *Advances in Neural Information Processing Systems*. M. C. Mozer, M. I. Jordan et T. Petsche (Eds.), Cambridge, MA, MIT Press. **9**, pp.473-479.

Hockett, C. (1966). What Algonquian is really like. *International Journal of American Linguistics,* 32: 59-73.

Hoeffner, J. et McClelland, J. L. (1993). *Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia ?A computational investigation*. Paper presented at the stanford Child LAnguage Research Forum, Stanford, CA.

Horne, B. G. et Giles, C. L. (1995). An experimental comparison of recurrent neural networks. *Advances in Neural Information Processing Systems*. G. Tesauro, D. Touretzky et T. Leen (Eds.), The MIT Press. **7**, pp. 697-704.

Houk, J. C. (1997). On the role of the cerebellum and basal ganglia in cognitive information processing. *Progress in Brain Research. The Cerebellum: From Structure to Control*. C. I. de Zeeuw, P. Strata et J. Voogd (Eds.), Amsterdam, Elsevier. **114**, pp.543-552.

Houk, J. C. et Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum and cerebral cortex: their role in planning and controlling action. *Cerebral Cortex,* 5: 95-110.

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R. et Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review,* 7: 504-509.

Houtsma, A. J. M. et Goldstein, J. L. (1972). The central origin of the pitch of complex tones: Evidence from musical interval recognition. *Journal of the Acoustical Society of America,* 51: 520-529.

Howell, P., Au-Yeung, J., Davis, S., Charles, N., Sackin, S., Williams, R., Cook, F., Rustin, L. et Reed, P. (1999). *Factors implicated in the diagnosis and prognosis of children who stutter*. The Oxford Dysfluency Conference 1999.

Hulse, S. H., Page, S. C. et Braaten, R. F. (1990). An integrative approach to auditory perception by songbirds. *Comparative perception, vol. 2 : complex signals*. W. C. stebbins et M. A. Berkley (Eds.), New York, Wiley & sons, pp.3-34.

Hung, F.-S. et Peters, A. M. (1997). The role of prosody in the acquisition of grammatical morphemes: evidence from two Chinese languages. *Journal of Child Language,* 24: 627-650.

Imberty, M. (1969). *L'acquisition des structures tonales chez l'enfant*. Paris, Klincksieck.

Ingram, D. (1974). The acquisition of English verbal auxiliary and copula in normal and linguistically deviant children. *Papers and reports on Child Language Developement 4* **:** 79-91.

Itahashi, S. et Du, L. (1995). *Language identification based on speech fundamental frequency*. In Eurospeech, volume 2, pp. 1359-1362.

Itahashi, S., Kiuchi, T. et Yamamoto, M. (1999). *Spoken Language Identification Utilizing Fundamental Frequency and Cepstra*. in Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99), Budapest, Hungary,.

Ivry, R. (1996). The representation of temporal information in perception and motor control. *Current Opinion in Neurobiology,* 6: 851-857.

Ivry, R. et Keele, S. W. (1989). Timing fonctions of the cerebellum. *Journal of Cognitive Neurosciences,* 1: 136-152.

Jescheniak, J. D., Hahne, A. et Friederici, A. D. (1998). Brain activity patterns suggest prosodic influences on syntactic parsing in the comprehension of spoken sentences. *Music perception,* 16: 55-62.

Joel, D., Niv, Y. et Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks,* 15: 535–547.

Johnson, E. K. et Jusczyk, P. W. (2001). Word segmentation by 8-month-olds : When speech cues count more than statistics. *Journal of Memory and Language,* 44: 1-20.

Johnsrude, J., Penhune, V. B. et Zatorre, R. J. (2000). Functional specificity in right human auditory cortex for perceiving pitch direction. *Brain,* 123: 155-163.

Jones, M. R., Summerell, L. et Marshburn, E. (1987). Recognizing melodies: A dynamic interpretation. *Quarterly Journal of Experimental Psychology,* 39A: 89-121.

Jonides, J., Smith, E. E., Koeppe, R. A., Awh, E., Minoshima, S. et Mintun, M. (1993). A Spatial working memory in humans as revealed by PET. *Nature,* 363: 623-625.

Jordan, M. I. (1986a). Serial order: A parallel distributed processing approach. *Technical Report ICS Report 8604, Institute for Cognitive Science, University of California at San Diego, La Jolla, CA.*

Jordan, M. I. (1986b). *Attractor dynamics and parallelism in a connectionist sequential machine.* In Proceedings of the Eighth Annual conference of the Cognitive Science Society, pages 531-546. Lawrence Erlbaum.

Jordan, M. I. (1990). Learning to articulate: Sequential networks and distal constraints. *Attention and Performance XIII*. M. Jeannerod (Ed.), Hillsdale , NY, Lawrence Erlbaum.

Jusczyk, P. W. (1989). *Perception of cues to clausal units in native and non-native languages*. in the Biennal Meeting of the Society for Research in Child Development, Kansas City.

Jusczyk, P. W. (1997). *The discovery of spoken language.*, MIT Press.

Jusczyk, P. W. (2002). L'apprentissage du langage : ce que le nourrisson sait et ce nous en ignorons. *Les langages du cerveau*. E. Dupoux (Ed.), Paris, Odile Jacob.

Jusczyk, P. W. et Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of Experimental Psychol Hum Percept Perform,* 19(3): 627-40.

Jusczyk, P. W. et Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology,* 29(1): 1-23.

Jusczyk, P. W. et Kemler Nelson, D. G. (1996). Syntatic units, prosody, and psychological reality during infancy. *Signal to syntax : Boostrapping from speech to grammar in early acquisition*. J. L. Morgan et K. Demuth (Eds.), pp.389-408.

Jusczyk, P. W., Pisoni, D. B. et Mullenix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition,* 43: 253-291.

Jusczyk, P. W., Cutler, A. et Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development,* 64(3): 675-87.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L., Woodward, A. et Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology,* 24: 252-293.

Karmarkar, U. R. et Buonomano, D. V. (1998). Temporal Specificity of Perceptual Learning in an Auditory Discrimination Task.

Kawahara, H., Katayose, H., de Cheveigné, A. et Patterson, R. D. (1999). *Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity*. Proc. EUROSPEECH 6, pp. 2781-2784.

Kelly, M. H. (1988). Phonological biases in grammatical category shifts. *Journal of Memory and Language,* 27: 343-358.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review,* 99: 349-364.

Kelly, M. (1996). The role of phonology in grammatical category assignment. *From Signal to Syntax*. J. Morgan et K. Demuth (Eds.), Hillsdale, Erlbaum, pp.249-262.

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W. et Wright, C. K. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language,* 16: 53-68.

Kermadi, I. et Joseph, J. P. (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *Journal of Neurophysiology,* 74: 911-933.

Kermadi, I., Jurquet, Y., Arzi, M. et Joseph, J. P. (1993). Neural activity in the caudate nucleus of monkeys during spatial sequencing. *Experimental Brain Research,* 94: 352-356.

Kilborn, K. (1991). Selective impairment of grammatical morphology due to induced stress normal listeners: Implications for aphasia. *Brain and Language,* 41: 275–288.

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: a critical review. *Psychological Bulletin,* 112: 24-38.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected

discourse. *Journal of Phonetics,* 3: 129-140.

Klatt, D. H. (1976). Linguistic uses os segmentatl duration in English. Acoustic and perceptual evidence. *Journal of the Acoustical Society of America,* 59: 1208-1221.

Klatt, D. K. (1980). Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access. *Production and Perception of Fluent Speech*. R. Cole (Ed.), Lawrence Erlbaum.

Klatt, D. H. et Cooper, W. E. (1975). Perception if segment duration in sentence contexts. *Cohen and Nooteboom*: 69-86.

Kluender, K. R., Lotto, A. J., Holt, L. L. et Bloedel, S. L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *Journal of Acoustic Society of America,* 104: 3568-3582.

Kohavi, Z. (1978). *Switching and Finite Automata Theory.* I. McGraw-Hill (Ed.), New York, NY.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* 43: 59-69.

Konopczynski, G. et Tessier, S. (1994). Structuration intonative du langage émergent. *Halford and Pilch,* 157-192. (cité dans dodane, 2003).

Kremer, S. C. (1995). *On the computational power of Elman-style recurrent networks.* IEEE Transactions on Neural Networks, 6(4).

Kremer, S. C. (2001). Spatiotemporal Connectionist Networks: A Taxonomy and Review. *Neural Computation,* 13: 249-306.

Kuhl, P. (1991). Human adults and human infants perceptual magnet effects. *Perception & Psychophysics,* 50: 93-157.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. et Lindblom, B. (1992). Linguistic experiences alter phonetic perception in infants by 6 months of age. *Science,* 255: 606-608.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U. et Lacerda, F. (1997). Cross-language analysis of phonetic units in language adressed to Infants. *Science,* 277: 684-686.

Kwasny, S., Kalman, B., Wu, W. et Engebretson, A. (1992). *Identifying Language from Speech: An Example of High-Level, Statistically-Based Feature Extraction*. In Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.

Lacheret-Dujour, A. et Morel, M. (2001). *Génération automatique de la prosodie pour la synthèse à partir du texte : le système KALI*. Journées Prosodie, pp. 57-60.

Ladd, D., Mennen, I. et Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America,* 107.

Ladefoged, P. (1975). *A course in phonetics*. New York, Harcourt Brace Jovanovich.

Lamel, L. F. et Gauvain, J. L. (1994). *Language Identification Using Phone-based Acoustic likelihoods,*. Proc. of ICASSP'94, pp. 293-296, Adalaida.

Lapicque, L. (1907). Recherches quantitatives sur l'excitation életrique des nerfs traitée comme une polarisation. *Journal de Neurophysiologie Paris,* 9: 620-635.

Large, E. W. et Kolen, J. F. (1994). Resonnance and the perception of musical meter. *Connection Science,* 6(1): 177-208.

Large, E. W. et Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science,* 26: 1-37.

Lashley, K. S. (1951). The problem of serial order in behavior. *Cerebral Mechanisms in Behavior*. L. A. Jeffres (Ed.), New York, Wiley, pp.112-136.

Laver, J. (1994). *Principles of phonetics*. Cambridge Textbooks in Linguistics, University Press.

Leavers, V. F. et Burley C.E. (2001). The use of cognitive processing strategies and linguistic cues for efficient automatic language identification. *Language Science,* 23: 639-650.

Lecanuet, J.-P. (1997). Dans tous les sens... bref état des compétences sensorielles fœtales. *Que savent les foetus ?*, Toulouse, E.R.E.S., pp.17-34.

Lee, L. (1966). Developmental sentences analysis. A method for comparing normal and deviant syntactic development. *Journal of speech and hearing disorders,* 31: 311-330.

Leman, M., Lesaffre, M. et Tanghe, K. (2001). Toolbox for perception-based music analysis : Concepts, demos, and reference manual, Institute for Psychoacoustics and Electronic Music (IPEM).

Lenormand, L., Leonard, L. et McGregor, K. (1993). A cross-linguistic study of article use by children with SLI. *European journal of disorders of communication,* 28: 153-163.

Leonard, L. (1985). Unusual and subtle phonological behavior in the speech of phonologically-disordered children. *Journal of Speech and Hearing Disorders,* 50: 4-13.

Leonard, L. (1989). Language learnability and specific language impairment in children. *Applied Psycholinguistics,* 10: 179-202.

Leonard, L. (1995). Functional categories in the grammars of children with specific language impairment. *Journal of Speech and Hearing Research,* 38: 1270-1283.

Leonard, L. (1998). *Children with specific language impairment*. Cambridge, MA, MIT Press.

Leonard, L., McGregor, K. et Allen, G. (1992). Grammatical morphology and speech perception in children with specific language impairment. *Journal of speech and hearing research,* 35: 1076-1085.

Lerdahl, F. et Jackendoff, R. S. (1985). *A Generative Theory of Tonal Music*. Cambridge, M.I.T. Press.

Lewicki, M. S. et Arthur, B. J. (1996). Hierarchical organization of auditory temporal context sensitivity. *Journal of Neuroscience,* 16: 6987-6998.

Li, K. P. et Edwards, T. J. (1994). *Automatic Language Identification Using Syllabic Spectral Features*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, pp. 297-300, Adelaide, Australia.

Liaw, J.-S. et Berger, T. W. (1996). Dynamic Synapse: A New Concept of Neural Representation and Computation. *Hippocampus,* 6: 591-600.

Liaw, J.-S. et Berger, T. W. (1998). *Robust speech recognition with dynamic synapses*. Proceedings of IJCNN'98, pp. 2175-2179.

Liaw, J.-S. et Berger, T. W. (2000). Dynamic synapse: Harnessing the computing power of synaptic dynamics. *Neurocomputing,* 26-27: 199-206.

Lieberman, P. (1965). On the acoustic basis of the perception of intonation by linguist. *Word,* 21: 40-54.

Liégeois-chauvel, C., De Graaf, J. B., Laguitton, V., C. et Hauvel, P. (1999). Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cerebral Cortex,* 9: 484-496.

Lin, T., Horne, B. G., Tiño, P., and Giles, C. L. (1996). *Learning long-term dependencies in NARX recurrent neural networks.* IEEE Transactions on Neural Networks, 7(6):1329-1338. 1424-1438.

Lively, S. E., Logan, J. S. et Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America,* 94: 1242-1255.

Llisterri (1996). Prosody tools efficiency and failures. *WP4 corpus T4.6 Speech markup and validation LRE-Project 62-050 MULTEXT*.

Locke, J. (1993). *The child's path to spoken language*. Cambridge, MA, Harvard University Press.

Lorch, M. P. et Meara, P. (1989). How people listen to languages they don't know. *Language Science,* 11(4): 343-353.

Lorch, M. P. et Meara, P. (1995). Can people discriminate language they don't know ? *Language Science,* 17(1): 65-71.

Lowe, A. et Campbell, R. N. (1965). Temporal discrimination in aphasoid and normal children. *Journal of Speech and Hearing Research,* 8: 313-314.

Lu, T. et Wang, X. (2000). Temporal Discharge Patterns Evoked by Rapid Sequences of Wide- and Narrowband Clicks in the Primary Auditory Cortex of Cat. *J Neurophysiol,* 84(1): 236-246.

Lu, T., Liang, L. et Wang, X., . (2001). Neural representations of temporally asymmetric stimuli in the auditory cortex of awake primates. *Journal of Neurophysiology,* 85: 2364-2380.

Luce, P. A. et Charles-Luce, J. (1983). *The role of fundamental frequency and duration in the perception of clause boundaries: Evidence from a speeded verification task*. the 105th meeting of the Acoustical Society of America, Cincinatti.

Ludlow, C. L., Cudahy, E. A., Bassich, C. et Brown, G. L. (1983). Auditory processing skills of hyperactive, language-impaired, and reading-disabled boys. *Central auditory processing disorders*. E. Z. Lasky et J. Katz (Eds.), Baltimore, MD, University Park Press, pp.163-184.

Maidment, J. A. (1976). Voice fundamental frequency characteristics as language differentiators. *Speech and hearing: Work in progress, University College London* **:** 74-93.

Maidment, J. A. (1983). Language recognition and prosody: further evidence. *Speech,*

hearing and language: Work in progress, University College London. 1**:** 133-141.

Malfrère, F., Dutoit, T. et Mertens, P. (1998). *Automatic prosody generation using suprasegmental unit selection*. Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis.

Maratsos, M. et Chalkley, M. A. (1980). The internal language of children's syntax. *Children's Language*. K. Nelson (Ed.), New York, Gardner Press. **2**, pp.127-214.

Marchman, V. A. (1993). Constraints on plasticity in a connectionist model of the English past tense. *Journal of Cognitive Neuroscience,* 5(2): 215-234.

Marcus, S. M. (1975). Perceptual centers. Cambridge, UK., King's college (unpublished).

Marcus, G. F., Vijayan, S.,, Bandi Rao, S. et Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science,* 283 (5398): 77–80.

Margoliash, D. (1983). Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience,* 3: 1039-1057.

Marshall, J. C. (2002). Cognition et neurosciences : où en étions nous ? *Les langages du cerveau*. E. Dupoux (Ed.), Edition Odile Jacob.

Martin, P. (2001). *ToBi : L'illusion scientifique ?* Journées Prosodie, pp. 109-112, Grenoble.

Martin, C. S., Mullennix, J. W., Pisoni, D. B. et Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: learning, memory and cognition,* 15: 676-684.

Maskara, A. et Noetzel, A. (1992). *Forced simple recurrent neural networks and grammatical inference*. Proc. of the Fifteenth Annual Conference of the Cognitive Science Society, pp. 420-425.

Massaro, D. W. (1987). Categorical partition: a fuzzy logical model of categorization behavior. *Categorical Perception: The Groundwork of Cognition*. S. Harnad (Ed.), Cambridge, MA, University Press.

McAdams, S. et Bigand, E. (1994). *Penser les sons: La psychologie cognitive de l'audition humaine,*. Paris, Presses Universitaires de France.

McCarthy, G., Blamire, A. M., Puce, A., Nobre, A. C., Bloch, G., Hyder, F., Goldman-Rakic, P. et Shulman, R. G. (1994). *Functional magnetic resonance imaging of human prefrontal cortex activation during a spatial working memory task*. Proc. Natl. Acad. Sci. USA 91: 8690-8694, .

McClelland, J. E. et Elman, J. L. (1986). Interactive Processes in Speech Perception: The TRACE Model. *Parallel Distributed Processing*. Rumelhart et McClelland (Eds.), MIT Press.

McClurkin, J. W., Optican, L. M., Richmond, B. J. et Gawne, T. J. (1991). Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science,* 253: 675-677.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M. et Stroeve, S. (2000). *Approaching automatic recognition of emotion from voice : a rough benchmark*. In Proceedings of ISCA workshop on Speech and Emotion., Belfast.

Mechler, R., Victor, J. D., Purpura, K. P. et Shapley, R. (1998). Robust temporal coding

of contrast by V1 neurons for transient but not for steady-state stimuli. *Journal of Neuroscience,* 18: 6583-6598.

Meegan, D., Aslin, R. N. et Jacobs, R. A. (2000). Motor timing learned without motor training. *Nature Neuroscience,* 3: 860-862.

Mehler, J. et Bertoncini, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development,* 4: 271-284.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., et Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition,* 29: 143-178.

Mehler, J. et Christophe, A. (1995). Maturation and learning of language during the first year of life. *The Cognitive Neurosciences*. M. S. Gazzaniga (Ed.), Bradford Books / MIT Press, pp.953-954.

Mehler, J., Dupoux, E., Nazzi, T., et Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. J. L. Morgan et K. Demuth (Eds.), Mahwah, NJ, Lawrence Erlbaum Associates, pp.101-116.

Mercado III, E., Myers, C. E. et Gluck, M. A. (2001). A computationnel model of mechanisms controlling experience-dependant reorganisztion of representational maps in auditory cortex. *Cognitive, Affective & Behavorial Neuroscience,* 1(1): 37-55.

Mertens, P., Auchlin, P., Goldman, J.-P. et Grobet, A. (2001). *L'intonation du discours : une implémentation par balises ; motifs and premiers résultats*. Journées Prosodie, pp. 93-96.

Merzenich, M. M., Jenkins, W.M., Johnston, P., Schreiner, C., Miller, S.L. et Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science,* 271: 77-81.

Miall, C. (1989). The storage of time intervals using oscillating neurons. *Neural Computation,* 1: 359-371.

Middlebrooks, J. C., Clock, A. E., Xu, L. et Green, D. M. (1994). A panoramic code for sound location by cortical neurons. *Science,* 264: 842-844.

Miller, J. L. et Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics,* 25: 457-465.

Mintz, T., Newport, E. et Bever, T. (1995). *Distributional regularities in speech to young children*. In Proceedings of NELS 25, 43-54.

Mody, M. (1993). *Bases of reading impairment in speech perception: A deficit rate of auditory processing or in phonological encoding ?* Thèse, City University of New York, New York.

Monaghan, P., Chater, N. et Christiansen, M. H. (2003). *Inequality between the classes: Phonological and distributional typicality as predictors of lexical processing*. Proceedings of the cognitive science society conference, Boston, MA.

Moon, C., Cooper, R. P. et Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development,* 16: 495-500.

Morehead, D. et Ingram, D. (1973). The de velopment of base syntax in normal and linguistically deveiant children. *Journal of Speech and hearing Research,* 16:

330-352.

Morgan, J. (1986). *From simple input to complex grammar*. Cambridge, MA, MIT Press.

Morgan, J. L. et Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ, Lawrence Erlbaum Associates.

Morgan, J., Shi, R. et Allopena, P. (1996). Perceptual bases of rudimentary grammatical categories. *From Signal to Syntax*. J. Morgan et K. Demuth (Eds.), Erlbaum, Hillsdale, pp.263-283.

Mori, K., Toba, N., Harada, T., Arai, T., Komatsu, M., Aoyagi, M. et Murahara, Y. (1999). *Human Language Identification with Reduced Spectral Information*. in Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99), Budapest, Hungary, September 1999.

Morlec, Y. (1997). *Génération multiparamétrique de la prosodie du français par apprentissage automatique*. Thèse, Institut National Polytechnique de Grenoble, spécialité sciences cognitives, Grenoble.

Morlec, Y., Bailly, G. et Aubergé, V. (2001). Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication,* 33: 357-371.

Morrongiello (1988). The development of auditory pattern perception skills. *Advances in infancy research*. C. Rovee-Collier (Ed.), Norwood, NJ, Ablex. **5**, pp.135-172.

Muthusamy, Y. K., Jain, N. et Cole, R. A. (1994). *Perceptual Benchmarks for Automatic Language Identification*. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, Adelaide, Australia, April 1994.

Muthusamy, Y. K., Berkling, K., Arai, T., Cole, R. et Barnard, E. (1993). *A Comparison Of Approaches to Automatic Language Identification Using Telephone Speech*. In Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech 93), Berlin, Germany, September.

Nachtschläger, T., Maass, W. et Zador, A. (2000). Effecient temporal processing with biologically realistic dynamic synapses.

Nagarajan, S. S., Blake, D.T., Wright, B.A., Byl, N. et Merzenich, M. M. (1998). Practice-related improvements in somatosensory interval discrimination are temporally specific but generalize across skin location, hemisphere, and modality. *Journal of Neuroscience,* 18: 1559-1570.

Näger, C., Storck, J. et Deco, G. (2002). Speech recognition with spiking neurons and dynamic synapses: a model motivated by the human auditory pathway. *Neurocomputing,* 44-46: 937-942.

Nakagawa, S., Ueda, Y. et Seino, T. (1992). *Speaker-independent, text-independent Language Identification by HMM*. Proceedings of the 1992 International Conference on Spoken Language Processing (ICLSP 92), pp. 1011-1014, Alberta, Canada.

Narmavar, H. H., Liaw, J.-S. et Berger, T. W. (2001). *A new dynamic synapse neural network for speech recognition*. Proceeding of the IEEE International Joint Conference on Neural Networks, pp. 2985-2990.

Navon, D. (1977). Forest before trees: the precedence of global feature in visual perception. *Cognitive Psychology,* 9: 353-383.

Nazzi, T., Bertoncini, J. et Mehler, J. (1998). Language discrimination by newborns:

towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance,,* 24(3): 756-766.

Nazzi, T. , Jusczyk, P. W. et Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language,* 43: 1-19.

Nespor, M. et Vogel, I. (1986). *Prosodic Phonology*. Dordrecht, Foris.

Nicholson, J., Takahashi, K. , Nakatsu, R. (1999). *Emotion recognition in speech using neural networks*. 6th International Conference on Neural Information Processing, pp. 495-501.

Nicholson, J., Takahashi, K. et Nakatsu, R. (2000). Emotion Recognition in Speech Using Neural Networks,. *Neural Computing & Applications,* 9: 290-296.

Niv, Y., Joel, D., Meilijson, I. et Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors. *Adaptive Behavior*.

Nowik-Stern, A., Clarkson, M. G., Morris, M. K. et Bakeman, R. (1998). *The effect of premature infants' speech preferences on mother-preterm interaction (Poster)*. 11th Biennial International Conference on Infant Studies, Atlanta, Georgia.

Nwe, T. L., Foo, S. W. et De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication,* 41(4): 603-623.

Ogiela, D. (1995). *Pronoun case errors in normally developing and specifically language impaired children*. Thèse, Purdue University, Master's thesis.

Ohala, J. J. et Gilbert, J. B. (1981). Listeners' ability to identify languages by their prosody. *Problèmes de Prosodie: vol. 2, Expérimentations, Modèles et Fonctions*. P. Leon et M. Rossi (Eds.), Paris, Didier, pp.123-131.

Olshausen, B. A. et O'Connor, K. N. (2002). A new window on sound. *Nature Neuroscience,* 5: 292-293.

Ostendorf, M. et Veilleux, N. M. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary locations. *Computational Linguistics,* 20(1): 27-54.

Oudeyer, P. Y. (2002). *Novel useful features and algorithms for the recognition of emotions in human speech*. Proceedings of the 1st International Conference on Prosody (Speech Prosody 2002), pp. 547-550.

Palmeri, T. J., Goldinger, S. D. et Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology:Learning, Memory and Cognition,* 19: 309-328.

Pearlmutter, B. A. (1988). Dynamic recurrent neural networks. *Technical Report CMU-CS-88-191, Carnegie Mellon University.*

Pearlmutter, B. A. (1995). *Gradient calculations for dynamic recurrent neural networks: A survey.*IEEE Transactions on Neural Networks, 6(5):1212-1228.

Pellegrino, F. (1998). *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques,*. Thèse, Université Paul Sabatier, Toulouse.

Pellegrino, F., Chauchat, J.-H., Rakotomalala, R. et Farinas, J. (2002). *Can*

Automatically Extracted Rhythmic Units Discriminate among Languages? Proceedings of the 1st International Conference on Prosody (Speech Prosody 2002), Aix-en-Provence.

Peña, M., Bonatti, L. L., Nespor, M. et Mehler, J. (2002). Signal-driven computations in speech processing. *Science,* 298: 604-607.

Peretz, I. (2000). *Music perception and recognition*. B. Rapp (Ed.), Hove, Psychology Press.

Peters, A. M. et Strömqvist, S. (1996). The role of prosody in the acquisition of grammatical morphemes. *Signal to Syntax: Bootstrapping from speech to grammar in early acquisition*. J. M. K. Demuth (Ed.), Mahwah, NJ, Lawrence Erlbaum Associates, pp.215-232.

Petrides, M. (1991). *Functional specialization within the dorsolateral frontal cortex for serial order memory*. Proc. R. Soc. Lond. B Biol. Sci. 246: 299-306.

Pierrehumbert, J. (1980). *The phonology and phonetics of english Intonation*. Thèse, PhD thesis, MIT. Published by the indidana University Linguistic Club.

Pinker, S. (1984). Language learnability and language development. *Cambridge, MA: Havard University Press.*

Platzack, C. (1990). A grammar without functional categories: a syntactic study of early Swedish child language. *Nordic Journal of Linguistics,* 13: 107-126.

Plunkett, K. et Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition,* 48: 21-69.

Poddar, P. et Unnikrishnan, K. (1991). *Nonlinear prediction of speech signals using memory neuron networks*. Neural Networks for Signal Processing: Proceedings of the 1991 IEEE Workshop, pp. 395-404, IEEE Press.

Pollack, J. (1989). Implications of recursive distributed representations. *Advances in Neural Information Processing Systems*. D. Touretzky (Ed.), San Mateo, CA, Morgan Kaufmann. **1**, pp.527-536.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence,* 46: 77-105.

Pollack, J. (1994). Chapter 4: Limits of connectionism. *Technical report, Department of Computer Science, Ohio State University.*

Povel (1981). Internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance,* 7: 3-18.

Pratt, V. (1978). Six Lectures on Dynamic Logic. *Tech. Report MIT/LCS/TM-117, MIT.*

Prieto, P., van Santen, J. et Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics,* 23(4): 429-451.

Prior, A. N. (1967). *Past, Present and Future*. Oxford, University Press.

Proverbio, A. M., Minniti A. et Zani, A. (1998). Electrophysiological evidence of perceptual precedence of global vs. local visual information. *Cognitive Brain Research,* 6: 321-334.

Prut, Y., Vaadia, E., Berman, H., Haalman, I., Solvin, H. et Abeles, H. (1998). Spatiotemporal structure of cortical activity: properties and behavioral relevance.

Journal of Neurophysiology, 79: 2857-2874.

Pye (1983). Mayan telegrapheses : Intonational determinants of inflectional development in Quiche Mayan. *Language,* 59: 583-604.

Radford, A. S. (1990). *Syntactic theory and the acquisition of English syntax*. Oxford, Blackwell.

Rakowski, A. I. (1990). Intonation variants of musical intervals in isolation and in musical contexts. *Psychology of Music,* 18: 60-72.

Ramus, F. (1999). *Rythme des langues et acquisition du langage*. Thèse, EHESS, Paris.

Ramus, F. (2002a). *Acoustic correlates of linguistic rhythm: Perspectives*. Proceedings of speech prosody 2002.

Ramus, F. (2002b). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition,* 2.

Ramus, F., Nespor, M. et Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition,* 73: 265-292.

Ramus, F., Hauser, M. D., Miller, C., Morris, D. et Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science,* 274: 349-351.

Ratner, N. B. (1984). Patterns of vowel modification in mother-child speech. *Journal of Child Language,* 11: 557-578.

Rauschecker, J. P., Tian, B. et Hauser, M. (1995). Processing complex sounds in the macaque nonprimary auditory cortex. *Science,* 268: 111-114.

Reali, F., Christiansen, M. H. et Monaghan, P. (2003). *Phonological and Distributional Cues in Syntax Acquisition: Scaling-Up the Connectionist Approach to Multiple-Cue Integration*. Cognitive Science 2003, Boston, MA.

Redington, M., Chater, N. et Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science,* 22: 435–469.

Richmond, B. J., Optican, L. M. et Spitzer, H. (1990). Temporal encoding of two-dimensional patterns by single units in primate visual cortex I Stimulus-response relations. *Journal of Neurophysiology,* 64: 351-368.

Rilliard, A. et Aubergé, V. (2001). *Mesure de l'intellegibilté de la démarcation prosodique*. Journées Prosodie, pp. 89-92.

Riquimaroux, R. (1994). Neuronal auditory scene analysis ? *Trans Tech Comm Psychol Physiol Acoust Soc Jpn (H) 28:1-8.*

Rispoli, M. (1994). Pronoun case overextensions and paradigm building. *Journal of Child Language,* 21: 157-172.

Rodet, X. et Doval, B. (1992). Maximum-likelihood harmonic matching for fundamental frequency estimation. *Journal of Acoustic Society of America,* 92: 2428-2429.

Rosen, S. et Manganari, E. (2001). Is there a relationship between speech and nonspeech auditory processing in children with dyslexia? *Journal of Speech, Language and Hearing Research,* 44(4): 720-36.

Sachs, M. B. et Young, E. D. (1979). Encoding of steady-state vowels in the auditory

nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America,* 66: 470-479.

Saffran, J. R., Aslin, R. N. et Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science,* 274: 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N. et Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition,* 70: 27-52.

Samson, S., Ehrlé, N. et Baulac, M. (2001). Cerebral substrates for musical temporal process. *Ann NY Acad sci,* 930: 166-178.

Saussure, F. (1916/1949). *Cours de Linguistique Generale*. Paris, Librairie Payot.

Schaffer, D. (1984). The role of intonation as a cue to topic management in conversation. *Journal of Phonetics,* 12: 327-344.

Schlesinger, I. M. (1971). Production of utterances and language acquisition. *The Ontogenesis of Grammar*. D. I. Slobin (Ed.), New York, pp.63-101.

Schultz, W. et Romo, R. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements. I. Preparatory activity in the anterior striatum. *Experimental Brain Research,* 91: 363-384.

Schwartz, J. et Tallal, P. (1980). Rate of acoustic change may underlie hemispheric specialization for speech perception. *Science,* 207: 1380-1381.

Scott, D. R. (1982). Duration as a cue to the perception of the phrase boundary. *Journal of the Acoustical Society of America,* 71: 996-1007.

Sejnowski, T. J. et Rosenberg, C. R. (1986). NETtalk: a parallel network that learns to read aloud. *Technical Report JHU/EECS-86/01, John Hopkins University Electrical Engineering and Computer Science.* as published in Neurocomputing.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MIT Press.

Selkirk, E. O. (1996). The prosodic structure of function words. *Signal to Syntax : Bootstrapping from speech to grammar in early acquisition*. J. Morgan et K. Demuth (Eds.), Mahwah, NJ, Lawrence Erlbaum Associates., pp.187-213.

Sereno, J. A. et Jongman, A. (1995). Acoustic correlates of form class (unpublished), Cornell University.

Shady, M., Jusczyk, P. W., et Gerken, L. A. (1998). *Infants' sensitivity to function morphemes*. In 23rd Annual Boston University Conference on Language Development. Boston, MA.

Shady, M. E. et Gerken, L. A. (1999). Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language,* 26: 1-13.

Shafer, V., Gerken, L. A., Shucard, J., et Shucard, D. (1992). *" The " and the brain: An electrophysiological study of infant's sensitivity to English function morphemes*. In the Boston University Conference on Language Development, Boston, MA.

Shafer, V., Shucard, D., Shucard, J., et Gerken, L. (1998). An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal Of Speech, Language, and Hearing Research,* 41(4): 874-86.

Shahidullah, S. et Hepper, P. G. (1992). Hearing in the Feotus: Prenatal Detection of

Deafness. *International Journal of Prenatal and Perinatal Studies,* 4(3/4): 235-240.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. et Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science,* 270: 303-304.

Shi, R., Morgan, J. L. et Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category asssignment : a cross linguistic perspective. *Journal of Child Language,* 25: 169-201.

Shi, R., Werker, J. F. et Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words,. *Cognition,* Volume 72(2): B11-B21.

Siegelmann, H., Horne, B. et Giles, C. (1997). *Computational capabilities of recurrent narx neural networks.* IEEE Trans. on Systems, Man and Cybernetics.

Silverman, K. et Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. J. Kingston et M. Beckman (Eds.), Cambridge, Cambridge University Press.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. et Hirschberg, J. (1992). *ToBI: A standard for labelling English prosody*. Proceedings of the 1992 International Conference on Spoken Language Processing, pp. 867-870.

Slaney, M. (1998). Auditory Toolbox, Interval Research Corporation.

Slaney, M. et McRoberts, G. (1998). *Baby ears : a recognition system for affective vocalization*. proceedings of ICASSP98.

Slaney, M. et McRoberts, G. (2003). Baby ears : a recognition system for affective vocalization. *Speech Communication,* 39: 367-384.

Smith, M. R., Cutler, A., Butterfield, S. et Nimmo-smith, I. (1989). The perception of rhythm and word boundaries in noise masked speech,. *Journal of speech and hearing research,* 32: 912-920.

Spring, D. R. et Dale, P. S. (1977). Discrimination of linguistic stress in early infancy. *Journal of Speech and Hearing Research,,* 20: 224-232.

Steinhauer, K., Alter, K. et Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience,* 2: 191-196.

Stockmal, V., Muljani, D. et Bond, Z. S. (1996). *Perceptual Features of Unknown Foreign Languages as Revealed by Multi-dimensional Scaling*. ICSLP 96 Fourth international conference on spoken language processing.

Stockmal, M., Moates, D. et Bond, Z. S. (1998). *same talker, different language*. in Proceedings 5th international Conference on Spoken Language Processing. Vol. 2, pp. 93-98, Sydney, Australia.

Stockmal, V., Moates, D. et Bond, Z. S. (2000). Same talker, different language. *Applied Psycholinguistics,* 21: 383-393.

Sun, X. (2002). *The Determination, Analysis, and Synthesis of Fundamental Frequency*. Thèse, Northwestern University, Evanston, Illiois.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks,* 15.

Suri, R. E. et Schultz, W. (1998). Dopamine-like reinforcement signal improves learning

of sequential movements by neural network. *Experimental Brain Research,* 121: 350-354.

Suri, R. E. et Schultz, W. (1999). A neural network model with dopaminelike reinforcement signal that learns a spatial delayed response task. *Neuroscience,* 91: 871-890.

Suri, R. E., Bargas, J. et Arbib, M. A. (2001). Modeling functions of striatal

dopamine modulation in learning and planning. *Neuroscience,* 103: 65–85.

Sutcliffe, P., A. et Bishop, D. V. M. (2002). *Age-related changes in frequency discrimination: Tone duration and Training can affect performance*. Temporal Integration in Speech Perception, pp. 89.

Sutton, R. (1988). Learning to predict by methods of temporal difference. *Machine Learning, 3, 9-44*.

Sutton, R. S. et Barto, A. G. (1998). Reinforcement learning: An introduction. *Cambridge,* MA: MIT Press, Bradford Books.

Swanson, L., Leonard, L. et Gandour, J. (1992). Vowel duration in mothers'speech to young children. *Journal of Speech and Hearing Research,* 35: 617-625.

Tallal, P. (1994). Temporal coding in the brain. *the perception of speech time is of the essence*. G. Buzsaki, R. Llinas, W. Singer, A. Berthoz et Y. Christen (Eds.), Berlin, Springer, pp.291-299.

Tallal, P. et Piercy, M. (1973a). Defects on non-verbal auditory perception in children with developmental aphasia. *Nature,* 241: 468-469.

Tallal, P. et Piercy, M. (1973b). Developmental aphasia : Impaired rate of non-verbal processing as a function of sensory modality. *Neuropsychologia,* 11: 389-398.

Tallal, P. et Piercy, M. (1975). Developmental aphasia: The perception of brief vowels and extended stop consonants. *Neuropsychologia,* 13: 69-74.

Tallal, P., Stark, R., Kallman, C. et Mellits, D. (1981). A reexamination of some nonverbal perceptual abilities of language-impaired and normal children as a function of age and sensory modality. *Journal of Speech and Hearing Research,* 24: 351–7.

Tallal, P. et Miller, S. (1993). Neurobiological Basis of Speech: A case for the Preeminence of Temporal Processing. *Temporal Information Processing in the Nervous System: Special Reference to Dyslexia and Dysphasia*. R. Fitch, P. Tallal, A. M. Galaburda, R. R. Llinas et C. von Euler (Eds.), Annals of the New York Academy of Sciences. **682**, pp.27-47.

Tallal, P., Stark, R. et Mellits, D. (1985). Identification of language impaired children on the basis of rapid perception and production skills. *Brain and Language,* 25: 314-322.

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S.S., Schreiner, C., Jenkins, W.M. et Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science,* 271: 81-84.

Tallal, P., Merzenich, M., Miller, S. et Jenkins, W. (1998). Language Learning Impairments: Integrating Basic Science, Technology and Remediation. *Experimental Brain Research,* 123: 210-219.

Tank, D. W. et Hopfield, J. J. (1987). *Neural computation by concentrating information in time*. Proc Natl Acad Sci USA, pp. 1896-1900.

Thorpe, L. A. et Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. *Developmental Psychology,* 25: 122-127.

Thorsen, N. G. (1980). A study of the perception of sentence intonation - Evidence from Danish. *Journal of the Acoustical Society of America,* 67(3): 1014-1030.

Thymé-Gobbel, A. E. et Hutchins, S. E. (1996). *On using prosodic cues in automatic language identification*. in Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA, October 1996.

Tomblin, J. B. et Quinn, M. (1983). The contribution of perceptual learning to performance on the repetition task. *Journal of speech and hearing research,* 26: 369-372.

Tomblin, J. B. et Buckwalter, P. R. (1994). Studies of genetics of specific language impairment. *Specific language impairments in children*. R. Watkins et M. Rice (Eds.), Baltimore, Paul H. Brookes, pp.17-34.

Traber, C. (1992). F0 generation with a database of natural F0 patterns and with a neural network. *Talking Machines: Theories, Models and Designs*. G. Bailly, C.Benoît et T. R. Sawallis (Eds.), Amsterdam, The Netherlands, Elsevier, pp.287-304.

Trehub, S. E. et Trainor, L. J. (1994). Les stratégies d'écoute chez le bébé : origines du développement de la musique et de la parole. *Penser les sons : Psychologie cognitive de l'audition*. S. McAdams et E. Bigand (Eds.), Paris, PUF.

Trehub, S. E. et Henderson, J. L. (1996). Temporal resolution in infancy and subsequent language development. *Journal of speech and hearing research,* 39: 1315-1320.

Treisman, M. (1963). Temporal discrimination and the indifference interval: implications for a model of the "internal clock". *Psychol Monogr,* 77: 1-31.

Ullman, M. T. et Pierpont, E. I. (In Press). Specific Language Impairment is not Specific to Language: The Procedural Deficit Hypothesis. *Cortex,*(The neuronbiology of developmental disorders edited by D. Bishop, M. Eckert and C. Leonard).

Van de Weijer, J. (2001). Vowels in infant- and adult-directed speech. *Lund University, Dept. Of linguistics working Papers 49*: 172-175.

Van der Lely, H. K. J. (1996). Specifically language impaired and normally developing children: Verbal passive vs. adjectival passive sentence interpretation. *Lingua,* 98(4): 243-272.

van Heuven, V. J., Haan, J., Janse, E. et vanDer Torre, E. J. (1997). *Perceptual identification of sentence type and the time-distribution of prosodic interrogativity marker in Dutch*. ESCA Workshop on Intonation: Theory, Models and Applications, pp. 317-320, Athens, Greece.

van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences.*Thèse, The Netherlands, Eindhoven University of Technology.

Vasilescu, I., Hombert, J.-M. et Pellegrino, F. (2000). *Détermination expérimentale d'indices perceptuels pour la différenciation des langues romanes*. Proc. of the 1st Freiburg Workshop on Romance Corpus Linguistics, Freiburg, Allemagne, 2000.

Veilleux, N. M., Ostendorf, M. et Wightman, C. W. (1992). *Parse Scoring with Prosodic Information*. Proc. 1992 Intl. Conf. on Spoken Language Processing, pp. 1605-1608.

Venditti, J. J., Jun S. et Beckman, M. (1996). Prosodic cues to syntactic and other linguistic structures in Japanese, Korean, and English. *Signal to syntax : Boostrapping from speech to grammar in early acquisition*. J. L. Morgan et K.Demuth. (Eds.).

Vendler, Z. (1957). Verbs and Times. *Philosophical Review,* 66: 143-160.

Voegtlin, T. (2002a). Recursive Self-Organizing Maps. *Neural Networks,* 15(8-9): 979-991.

Voegtlin, T. (2002b). *Neural Networks and Self-Reference.*Thèse, Universite Lyon 2.

Waibel, A. (1988). Consonant recognition by modular construction of large phonemic time-delay neural networks. *Neural formation Processing Systems*. D. Anderson (Ed.), New York, NY, American Institute of Physics, pp.215-223.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). *Phoneme recognition using time-delay neural networks.* IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 328-339.

Wang, X. (2000). *On cortical coding of vocal communication sounds in primates*. Proc Natl Acad Sci USA 97: 11843-11849.

Wang, L. et Alkon, D. L. (1995). An artificial neural network system for temporal-spatial sequence processing. *Pattern Recognition,* 28(8): 1267-1276.

Wang, X., Merzenich, M. M., Beitel, R. et Schreiner, C. E. (1995). Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *Journal of Neurophysiology,* 74: 2685-2706.

Warren, R. M. (1994). La perception des séquences acoustiques : intégration globale ou résolution temporelle ? *Penser les sons : Psychologie cognitive de l'audition.* S. McAdams et E. Bigand (Eds.), Paris, PUF.

Warren, R. M. et Byrnes, D. L. (1975). Temporal discrimination of recycled tonal sequences : Pattern matching and naming of order by untrained listeners. *Perception and Psychophysics,* 12: 86-90.

Warren, R. M. et Ackroff, J. M. (1976). Two types of auditory sequences perception. *Perception and Psychophysics,* 20: 387-394.

Watson, C. S. et Kelly, W. J. (1981). The role of stimulus uncertainty in the discrimination of auditory patterns. *Auditory processing of complex sounds*. W. A. Yost et C. S. Watson (Eds.), Hillsdale, NJ, L. Erlbaum Assoc., pp.267-277.

Weber, J. (1989). *A Parallel Algorithm for Statistical Refinement and its use in Causal reasoning*. Proc. International Joint Conference on Artificial Intelligence (IJCAI89), Morgan Kaufman.

Westheimer, G. (1999). Discrimination of short time intervals by the human observer. *Experimental Brain Research,* 129: 121-126.

Wetzel, W., Ohl, F. W., Wagner, T. et Scheich, H. (1998). Right auditory cortex lesion in Mongolian gerbils impairs discrimination of rising and falling frequency-modulated tones. *Neuroscience Letters,* 252(2): 115-118.

Williams, R. J. et Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation,* 1(2): 270-280.

Wright, B. A., Buonomano, D. V., Mahncke, H. W. et Merzenich, M. M. (1997b). Learning and generalization of auditory temporal-interval discrimination in humans. *Journal of Neuroscience,* 17: 3956-3963.

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M. et Merzenich, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature,* 387(6629): 176-178.

Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica,* 58: 26-52.

Yan, Y. et Barnard, E. (1995). *An Approach to Language Identification with Enhanced Language Model*. Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech 95), Madrid, Spain, September.

Young, E. D. et Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. *Journal of the Acoustical Society of America,* 66: 1381-1403.

Zatorre, R. J. (1988). Pitch perception of complex tone and human temporal-lobe fonction. *Journal of the Acoustical Society of America,* 84: 566-572.

Zatorre, R. J. et Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex,* 11 (10): 946-953.

Zatorre, R. J., Evans, A. C., Meyer, E. et Gjedde, A. (1992). Lateralization of phonetic and pitch processing in speech perception. *Science,* 256 (5058): 846-849.

Zellner Keller, B. (2001). *Les enjeux de la simulation scientifique,. L'exemple du rythme de la parole*. Journées Prosodie, pp. 99-102, Grenoble.

Zhang, X. et Tomblin, J. B. (1999). Can children with language impairment be accurately identified using temporal processing measures? A simulation study. *Brain and Language,* 65: 395-403.

Zissman, M. A. (1993). *Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 93, Vol.2, pp. 399-402, Minneapolis, USA.

Zissman, M. A. (1996). *Comparison of Four Approaches to Automatic Language Identification of Telephone Speech*. IEEE Trans. Speech and Audio Proc., SAP-4(1), January.

Zissman, M. A. et Martin, A. (1995). *Language Identification Overview*. Proceedings of 15th annual speech research symposium Johns Hopkins University no.1, pp. 31-44.

Zissman, M. A. et Singer, E. (1995). *Language Identification Using Phoneme Recognition and Phonotactic Language Modeling*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 3503-3506.

Zissman, M. A. et Singer, E. (1995). *Language Identification Using Phoneme Recognition and Phonotactic Language Modeling*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 3503-3506.