REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT D'ELECTRONIQUE

PROJET DE FIN D'ETUDES

EN VUE DE L'OBTENTION DU DIPLOME
D'INGENIEUR D'ETAT EN ELECTRONIQUE

THEME

MISE EN ŒUVRE DE LA COMPOSANTE LEXICO-SYNTAXIQUE DANS UN SYSTEME DE RECONNAISSANCE

Proposé et dirigé par :

-M. BOUSSEKSOU

Réalisé par :

-M. KACETE Amine

-M.OUKSILI Youcef

Promotion : Juin 2011

ملخص

يندرج هذا العمل ضمن إطار تعرف الأوتوماتيكي لكلام،وإساس تعمال عناصر الالنحو و الصرف. هذا الميدان الغني بالتطبيقات البالغة الأهمية بدءا من تامين المعابر و التطبيقات القضائية إلى تنظيم الملفات الصوتية. حتى نترك المجال مفتوح، اهتممنا بإطار تعرف الأوتوماتيك لكلام المرتبط بالنص المنطوق. اهتممنا بكيفية تمثيل الكلمات و استخراج الشارات الصوتية لكلمات حتى نتمكن من تطبيقها في نظام التعرف الكلمات لتنضيم النحوية القواعد بعض تصريح و عليها المتعرف

كلمات مفتاحية : المعزولة الكلمات على التعرف ، وسائط مال سابستر ، المخفية ماركوف نماذج ،النحو الصرف..

Résumé

Ce travail s'inscrit dans le domaine de la reconnaissance automatique de la parole et l'utilisation de la composante lexico-syntaxique, domaine riche d'applications potentielles allant de la traduction automatique à l'indexation de documents audio. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à l'identification des mots en mode dépendant du texte. Nous nous intéressons, plus particulièrement, à la modélisation et à la représentation des mots. Il s'agit d'extraire, à partir des signaux de parole, les informations relatives à chaque mot et d'estimer un modèle du mot permettant sa reconnaissance et définir certaines grammaires pour le module linguistique selon lesquelles ces mots doivent être organisés.

Mots clés : Reconnaissance de mots isolés, MFCC, modèles de Markov cachés HMM, lexique, syntaxe.

Abstract

This work relates to the automatic speech recognition and the use of the component lexicon-syntax which has many potential applications ranging from automatic translation to audio indexing. In this thesis, the speech recognition is studied with a specific focus on word modeling and representation. We are especially interested to extract, from speech signals, the relative information of the word and estimate a sufficiently robust word's model and define grammars for the linguistic unit which this words must fellow.

Keywords: isolated word recognition, Mel frequency cepstral coefficients MFCC, Hidden Markov Model HMM, Lexicon, Syntax.

Remerciements

Nous tenons à remercier en premier lieu « ALLAH » le tout puissant, qui nous a donné la force, le courage et la volonté pour mener à bien ce modeste travail.

Nous exprimons notre profonde gratitude, notre grand respect et notre sincère reconnaissance à notre promoteur monsieur B. BOUSSEKSOU pour avoir assumé la lourde responsabilité de nous encadrer, de nous avoir orienté et conseiller tout au long de ce travail ainsi pour la confiance qu'il nous a accordée.

Nous remercions l'ensemble de nos enseignants d'Electronique de l'Ecole Nationale Polytechnique.

Nous remercions vivement tous nos enseignants et encadreurs de l'Ecole Nationale Préparatoire aux Etudes d'Ingéniorat.

Finalement, Nous remercions toute personne qui nous a soutenus de près ou de loin tout au long de notre parcours pour la réalisation de ce travail.

Dédicaces

A la mémoire de mes très chers grands-parents

A mes très chers parents.

A mon frère Bylkacem.

A mes sœurs Hayet, Farida et Dounia.

A ma tante Ouiza et son fils Zakaria.

A tous mes amis.

Amine

Dédicaces

A mes très chers parents.

A mes frères et sœurs.

A mes deux petites nièces

A mes grands parents.

A tous mes amis.

Youcef

Tables des matières

Table des figures

Liste des tableaux

Acronymes

AR : **A**uto **R**égressif

DCT: **D**iscrete **C**osine **T**ransform.

DTW: **D**ynamic **T**ime **W**arping.

EM: **E**xpectation **M**aximisation.

FFT: **F**ast **F**ourier **T**ransform.

HMM: **H**idden **M**arkov **M**odels.

LFCC: **L**inear **F**requency **C**epstral **C**oefficients.

LPC: **L**inear **P**rediction **C**oefficients.

LPCC: **L**inear **P**rediction **C**epstral **C**oefficients.

LSP: **L**ine **S**pectral **P**air (**L**ine **S**pectral **F**requencies)

MFCC: **M**el **F**requency **C**epstral **C**oefficients.

TF: **T**ransformée de **F**ourier

TFD: **T**ransformée de **F**ourier **D**iscrète.

PLP**: P**erceptual **L**inear **P**redictive.

GV : **G**roupe **V**erbal.

GN : **G**roupe **N**ominal.

DET : **D**éterminant.

VER : **V**erbe.

ADJ : **A**djectif.

CONJ : **C**onjonction.

PREP : **P**réposition.

Introduction générale

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Situé au croisement du traitement de signal numérique et du langage. Cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunication.

L'importance particulière du traitement de la parole dans ce cadre général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant dont dispose le cerveau humain a la fois dans la production et la compréhension de la parole et a l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée.

La parole est en effet produite par le conduit vocal, contrôlée en permanence par le cortex moteur. L'étude des mécanismes de phonation permettra donc de déterminer, dans une certaine mesure, ce qui est parole et ce qui n'en est pas. De même, l'étude des mécanismes d'audition et des propriétés perceptuelles qui s'y rattachent permettra de dire ce qui, dans le signal parole, est réellement perçu. Mais l'essence du signal de parole ne peut être cernée de façon réaliste que dans la mesure où l'on imagine, bien au-delà de la simple mise en commun des propriétés de production et de perception de la parole, les propriétés du signal dues a la mise en boucle de ces deux fonctions.

Mieux encore, c'est non seulement la perception de la parole qui vient influer sur sa production par le biais de ce bouclage, mais aussi et surtout sa compréhension. On ne parle

que dans la mesure ou l'on s'entend et ou l'on se comprend soi-même, la complexité du signal qui en résulte s'en ressent forcément.

CHAPITRE I

Production de la parole

La parole apparait physiquement comme une variation de pression de l'air causée et émise par le système articulatoire. Ce signal mécanique est transformé en signal électrique grâce à un transducteur approprié (microphone).

Ce signal électrique résultant peut être numérisé et soumis à un ensemble de traitements statistiques afin de mettre en évidence les *traits acoustiques*, sa *fréquence fondamentale*, son *énergie*, et *son spectre.* Chaque trait acoustique est lui-même intiment lié a une grandeur perceptuelle (pitch, intensité, timbre).

I.1 Physiologie des organes de phonation

Trois groupes d'organes assument les fonctions essentielles dans l'acte de parole, ou phonation :

- ✓ L'appareil respiratoire (diaphragme, poumons, trachées), soufflerie qui fournit l'énergie et la quantité d'air nécessaire.
- ✓ le larynx, organe vibrant, où naît le son.

- ✓ Le conduit vocal, formé des cavités résonantes supra laryngées (pharynx, bouche, nez)

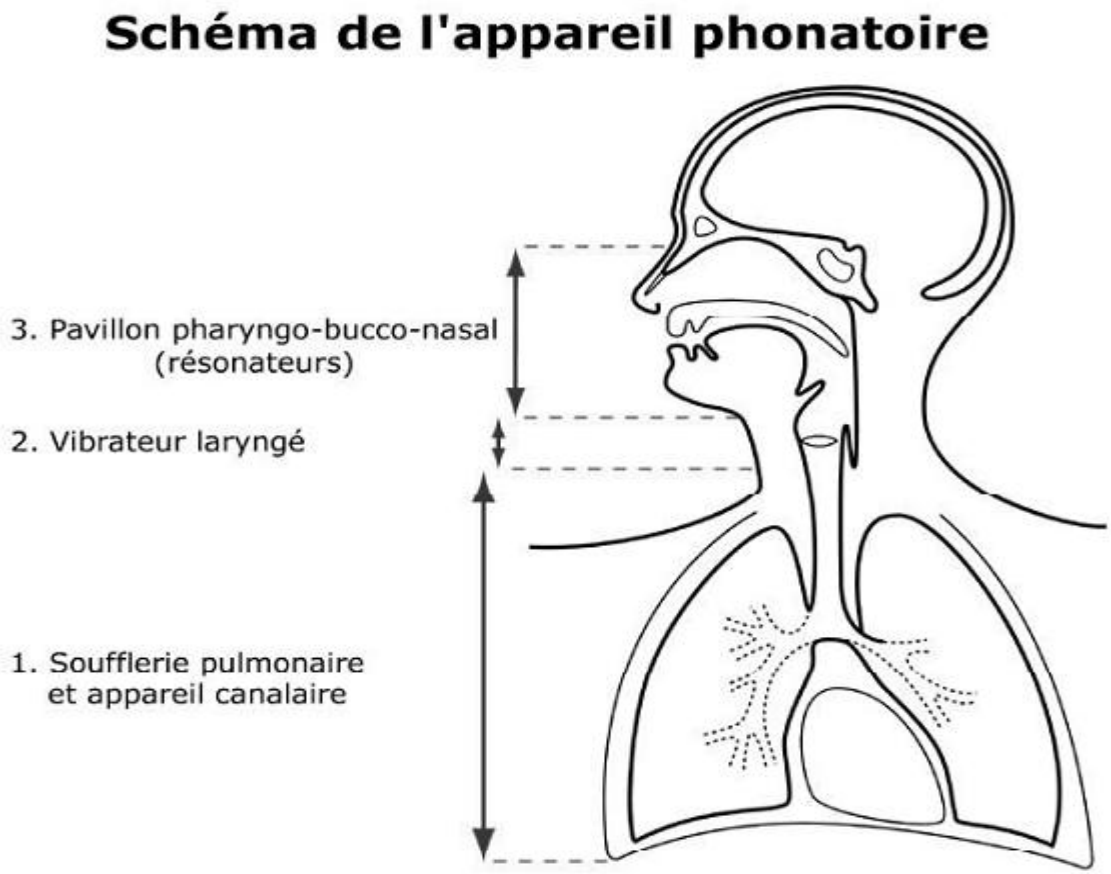


Figure I.1 L'appareil phonatoire [21]

L'appareil phonatoire humain schématisé :

- ✓ Partie sub glottique ou appareil respiratoire (diaphragme poumons, trachée) qui fournit l'énergie nécessaire à la phonation en insufflant l'air vers la partie glottique.
- ✓ Partie glottique ou larynx (ensemble de cartilages ligaments et muscles) contenant les cordes vocales (replis tendus horizontalement qui, sous l'effet des muscles, jouent un rôle de valve vis-à-vis de l'air des poumons libérant ainsi un flux d'air vers la partie supra glottique).
- ✓ Partie supra glottique ou conduit vocal, formé des cavités orales (pharyngienne et buccale), à géométrie variable, en fonction des éléments articulateurs (langue, mâchoire inférieure, lèvres) et des cavités nasales, à géométrie fixe, pouvant être couplées aux cavités orales par abaissement du voile du palais.

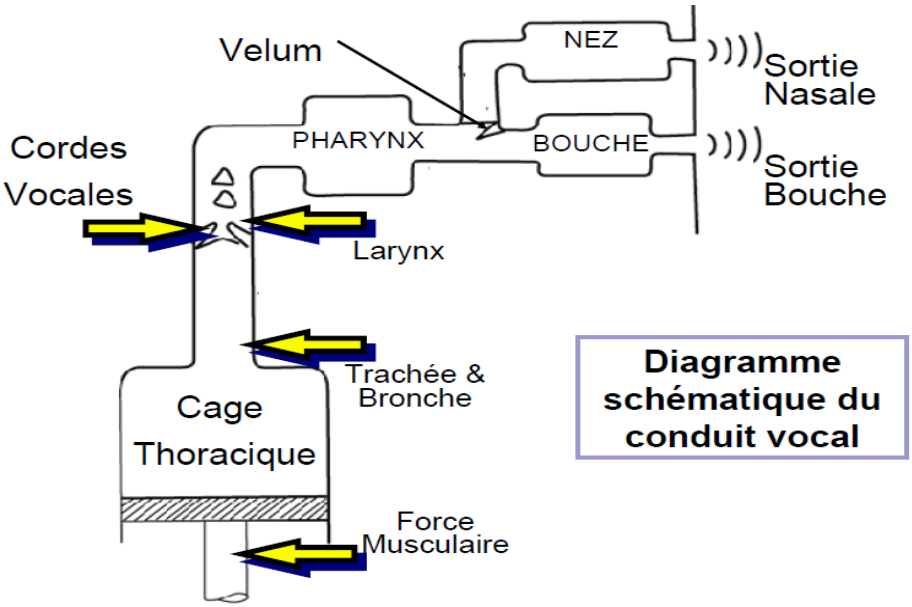


Figure I.2 Diagramme schématique du conduit vocal [21]

I.1.1 Le larynx, organe vibrant :

Organe de la phonation, puisqu'il joue un rôle très important dans l'émission des sons vocaux, le larynx est placé dans le cou à l'extrémité supérieure de l'arbre respiratoire.

Cet instrument vibrant est placé dans le cou, sur le trajet de l'air respiratoire, entre la soufflerie qui commande l'expiration, à savoir les poumons, et les cavités de résonance qui moduleront le son laryngé primaire.

Le larynx n'est pas fixe dans le cou il se déplace de haut en bas quand on parle. Il s'élève pour les sons aigus et s'abaisse pour les sons graves.

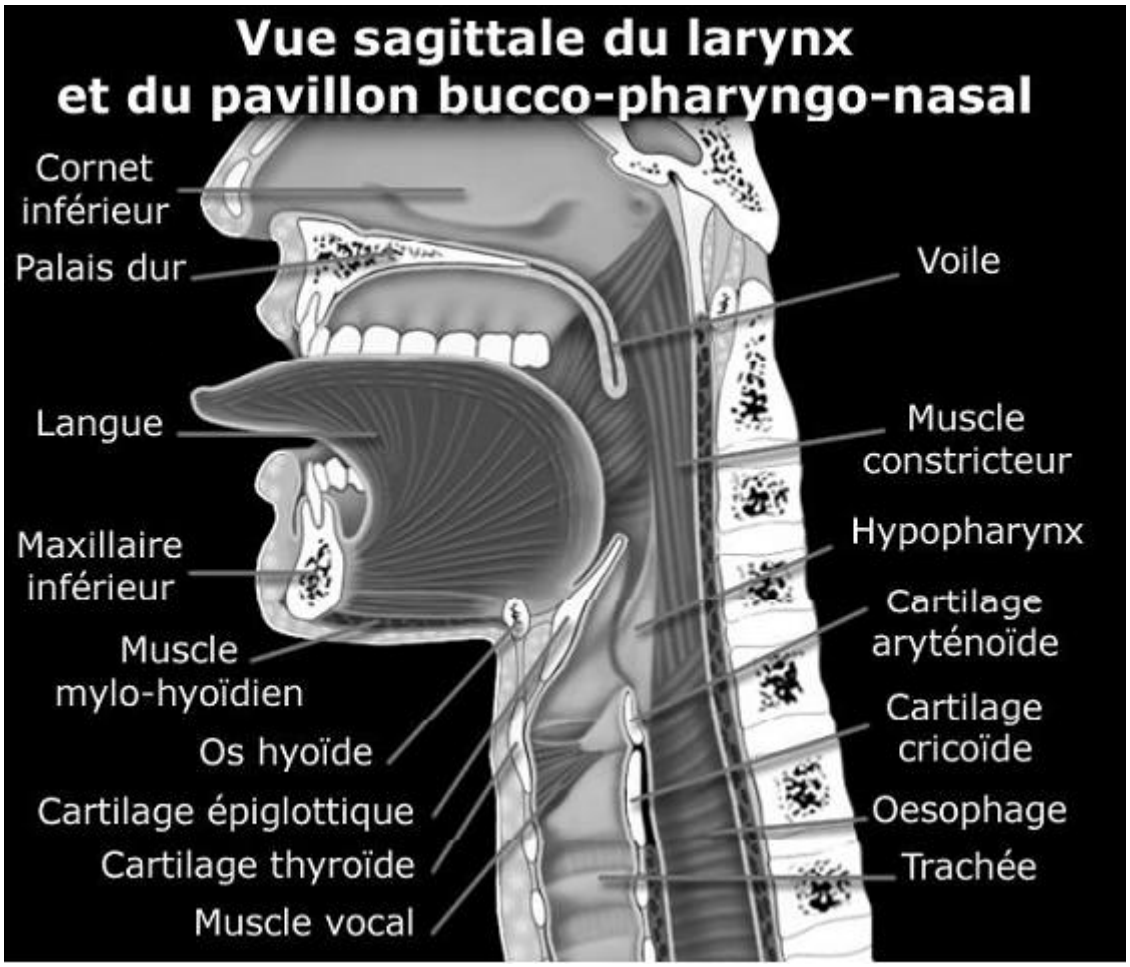


Figure I.3 Le larynx [21]

I.1.2 Les cordes vocales:

Une corde vocale est la superposition de deux muscles et d'un ligament.

Il y a tout d'abord, pour chaque corde vocale, un ligament vocal qui va du cartilage thyroïde à un cartilage aryténoïde.

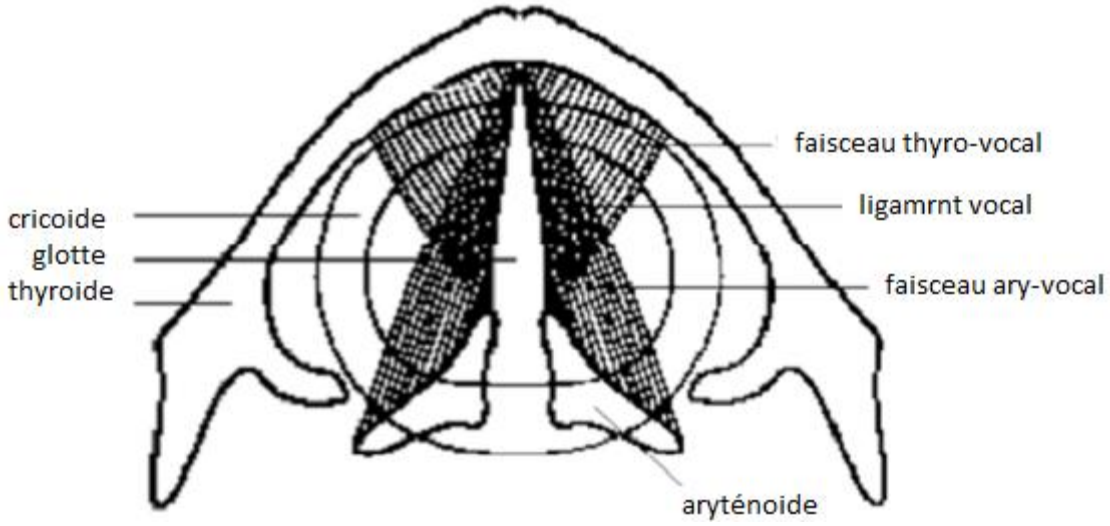
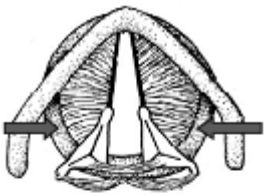
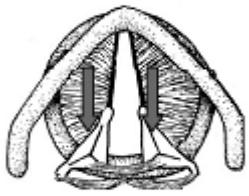


Figure I.4 Structures des cordes vocales

Les cordes vocales sont donc formées par des ligaments vocaux, longés par des muscles sur lesquels on peut agir; le tout est recouvert d'une muqueuse qui constitue la partie vibrante qui produit le son. Elles forment un clapet qui peut être ouvert ou fermé.



le cartilage cricoïde, en forme d'anneau



sur le quel sont posées deux pyramides, les cartilages aryténoïdes

et enfin, enserré comme un livre ouvert vers l'avant : le cartilage tyroïde

Figure I.5 Section schématique du larynx au niveau des cordes vocales

Le rôle des cordes vocales :

Les cordes vocales sont donc tendues de l'angle rentrant du cartilage thyroïde à l'apophyse vocale des cartilages aryténoïdes.

La glotte est l'espacement plus ou moins grand entre les deux cordes vocales.

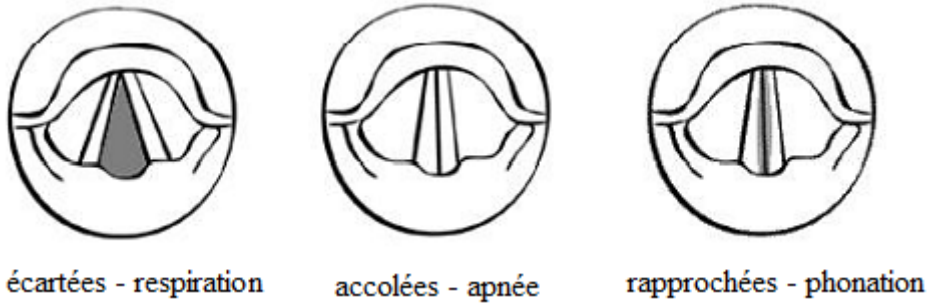


Figure I.6 L'espacement entre deux cordes vocales [21]

Les cordes vocales ont trois positions fondamentales :

- ✓ Soit, elles sont écartées : la glotte est ouverte et l'air circule librement. C'est la respiration. Lors d'une inspiration profonde, l'écartement est maximal, lors de la respiration normale, l'écartement est moyen.
- ✓ Soit, elles sont accolées : la glotte est alors fermée et l'air ne passe pas. C'est l'apnée.
- ✓ Soit, les cordes sont rapprochées : la glotte est variable. C'est la phonation ou le voisement.

I.1.3 Les cavités résonnantes

La majorité des sons du langage sont le fait du passage d'une colonne d'air venant des poumons, qui traverse un ou plusieurs résonateurs de l'appareil phonatoire. Les résonateurs principaux sont :

- ✓ le pharynx (ou cavité pharyngale) est un conduit musculo-membraneux situé entre la bouche et l'œsophage d'une part et entre les fosses nasales et le larynx d'autre part. La paroi du pharynx est constituée de muscles constricteurs. Effet d'une constriction :

modification du diamètre du pharynx. La racine de la langue peut également reculer ou avancer et donc agir sur le volume de cette première cavité supra glottique.

- ✓ les fosses nasales (ou cavités nasales) sont deux cavités cunéiformes séparées par une cloison verticale médiane et sont recouvertes de muqueuses. Une résonance nasale est très caractéristique (nasillement). L'air passe par le nez lorsque le voile du palais (prolongement) musculaire du palais osseux) est rabaissé : passage oro-nasal ouvert.

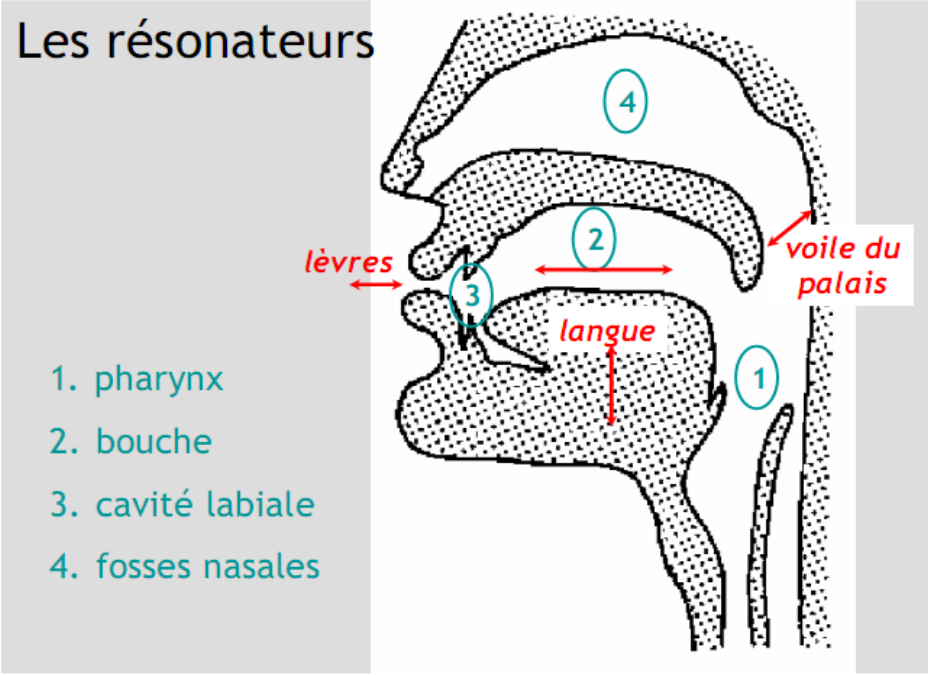


Figure I.7 Les résonateurs [21]

- ✓ la bouche (ou cavité buccale) est séparée des fosses nasales par une cloison appelée le palais. Dans cette cavité se situent des articulateurs, certains fixes (=passifs), d'autres mobiles (= actifs).

- ✓ la cavité labiale est une cavité que l'on crée lorsqu'on projette en avant les lèvres.

I.2 Production du son par l'appareil phonatoire

I.2.1 Fonctionnement général de l'appareil phonatoire

La parole est une succession d'évènements sonores faisant alternativement apparaître des sons dits voisés caractérisés par la vibration des cordes vocales et des sons non voisés (qui ne font pas intervenir les cordes vocales).

Le signal voisé est un signal pseudopériodique présentant des zones fréquentielles plus ou moins importantes. Ces zones fréquentielles d'enveloppe maximale sont appelées des formants.

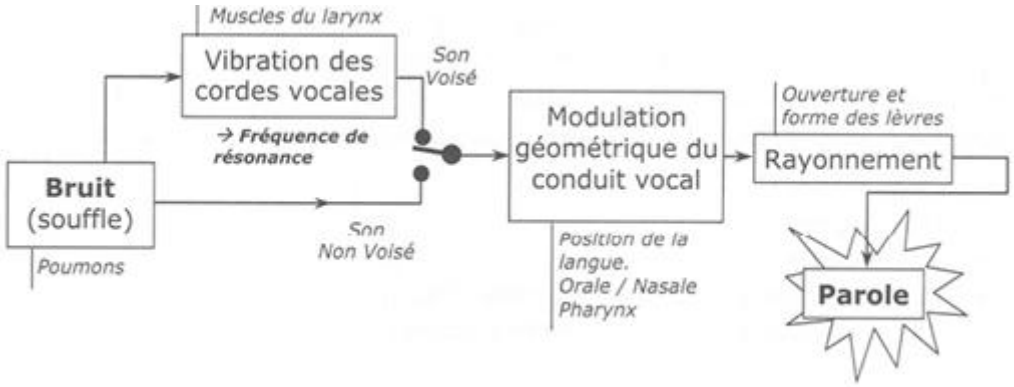


Figure I.8 Fonctionnement général de l'appareil phonatoire [21]

I.2.2 Mécanisme de production du son

Étape 1 : formation du flux d'air

Ce rôle est repris par les organes sub-glottaux. En particulier les poumons, mis en action par le diaphragme.

Celui-ci se contracte ce qui chasse l'air des poumons et c'est cet air dont les variations de pression au niveau du pharynx vont créer le son.

Étape 2 : création du son (au niveau des cordes vocales) ou phonation

Première sous étape : L'air expulsé des poumons arrive au niveau des cordes vocales au repos. Les cordes sont fermées (phase d'apnée) L'air s'y accumule donc jusqu'à ce que la pression exercée par la colonne d'air ne soit trop forte.

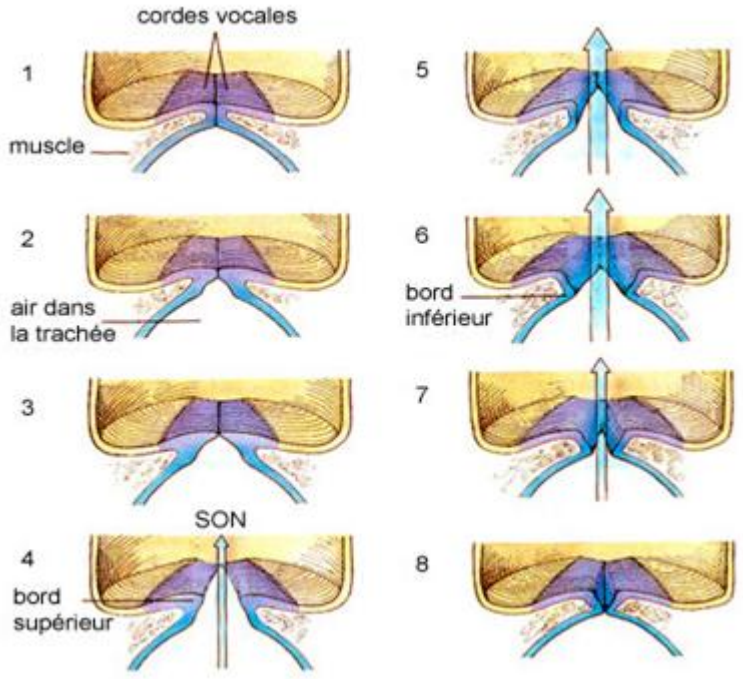


Figure I.9 Schéma représentant les cordes vocales et la tranchée [21]

- ✓ Deuxième sous étape : La pression d'air exercée sur le cordes vocales est suffisamment élevée pour les écarter (4). L'air commence à s'échapper par l'ouverture (5).La vitesse de ce fluide augmente alors, entrainant une chute de la pression sub glottique par le principe de Bernoulli. La partie inférieure des cordes vocales commence alors à se fermer (6).

- ✓ Troisième sous étape : La fermeture des cordes vocales se fait de manière évolutive. Leur structure est en effet assez complexe. La partie inférieure se ferme presque immédiatement. Cette fermeture provoque une chute brutale de la pression au niveau supérieur, qui se ferme en un claquement (7) et (8). L'air va donc à nouveau s'accumuler sous les cordes vocales, et le processus recommence.

I.2.3 Mise en forme du son

Cette étape correspond à l'amplification et au filtrage fréquentiel de l'onde source produite par la glotte.

Ces fonctions sont reprises de manière complexe par les organes supra glottiques.

On peut regrouper ces organes en quatre résonateurs qui renforcent ou atténuent certaines fréquences.

Ces résonateurs sont des cavités de forme et de taille variable, ce qui permet d'ajuster le timbre du son, via le phénomène de résonance.

Rappels physiques sur le phénomène de résonance :

Rappelons que le phénomène de résonance consiste en une modification du timbre du son résultant de l'enrichissement de certains de ses harmoniques et de l'appauvrissement pour d'autres.

Toute cavité, de par sa forme, présente une fréquence caractéristique ou fréquence propre : ce mode propre peut être excité par un son fondamental ou une de ses harmoniques, et la cavité résonne alors sous l'influence de ce son et le renforce.

La fréquence caractéristique varie selon la forme et le volume des cavités, les plus grandes ayant en général un son propre plus grave que les plus petites. Il en résulte qu'en aucun cas une cavité ne crée d'harmoniques nouveaux.

Des modifications de forme et de volume permettent donc un accord avec des sons de fréquences très diverses : les articulateurs façonnent dans le conduit vocal des cavités dont les fréquences de résonance déterminent la forme du spectre émis.

De plus, deux cavités résonantes ouvertes l'une sur l'autre modifient leurs propriétés, réciproques notamment en modifiant leur fréquence caractéristique, la plus grave s'aggravant et la plus aiguë devient encore plus aiguë. On parle d'un phénomène de couplage.

Résonances dans l'appareil phonatoire :

Le pharynx ne change pas facilement de forme, mais sa longueur peut changer légèrement en haussant ou en abaissant le larynx d'un côté ou le voile du palais de l'autre côté. Ce dernier agit aussi comme une valve qui permet d'isoler ou de connecter la cavité nasale au pharynx.

L'épiglotte agit aussi comme une valve, dont le rôle est d'empêcher toute nourriture d'atteindre le larynx : elle est ouverte au cours de la respiration normale mais se ferme au moment de la déglutition.

La cavité nasale a aussi des dimensions et une forme fixe. Pour un homme adulte, elle a une longueur d'environ 12 cm et un volume de l'ordre de 60 cm^3.

La cavité orale, ou buccale, est la partie la plus importante du tractus vocal car sa forme et sa taille peuvent varier en ajustant les positions relatives du palais, de la langue, des lèvres, et des dents.

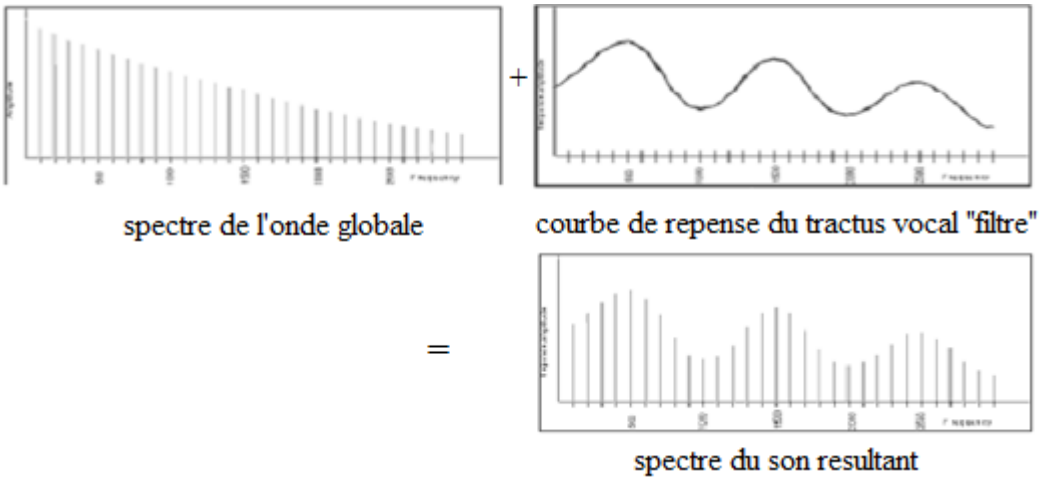


Figure I.10 Effets de la résonance sur le spectre émis

I.3.Notion de phonétique

En linguistique, un phonème est la plus petite unité distinctive (c'est-à-dire permettant de distinguer des mots les uns des autres) que l'on puisse isoler dans la chaîne parlée.

La phonétique traditionnelle classe les phonèmes en voyelles consonnes et semi voyelles (ou semi consonnes).

La distinction entre voyelles et consonnes s'effectue de la manière suivante :

- ✓ si le passage de l'air se fait librement à partir de la glotte, on a faire à une voyelle.
- ✓ si le passage de l'air à partir de la glotte est obstrué, complètement ou partiellement, en un ou plusieurs endroits, on a affaire à une consonne.

Les semi voyelles présentent la même articulation que les voyelles, mais se comportent dans la syllabe comme les consonnes : plus précisément, les consonnes et les semi voyelles ne peuvent constituer à elles seules une syllabe, les voyelles si : par exemple, le mot abbaye [a /be / i] comporte des voyelles alors que le mot abeille [a / bɛj] comporte aussi une semi-voyelle notée [j].

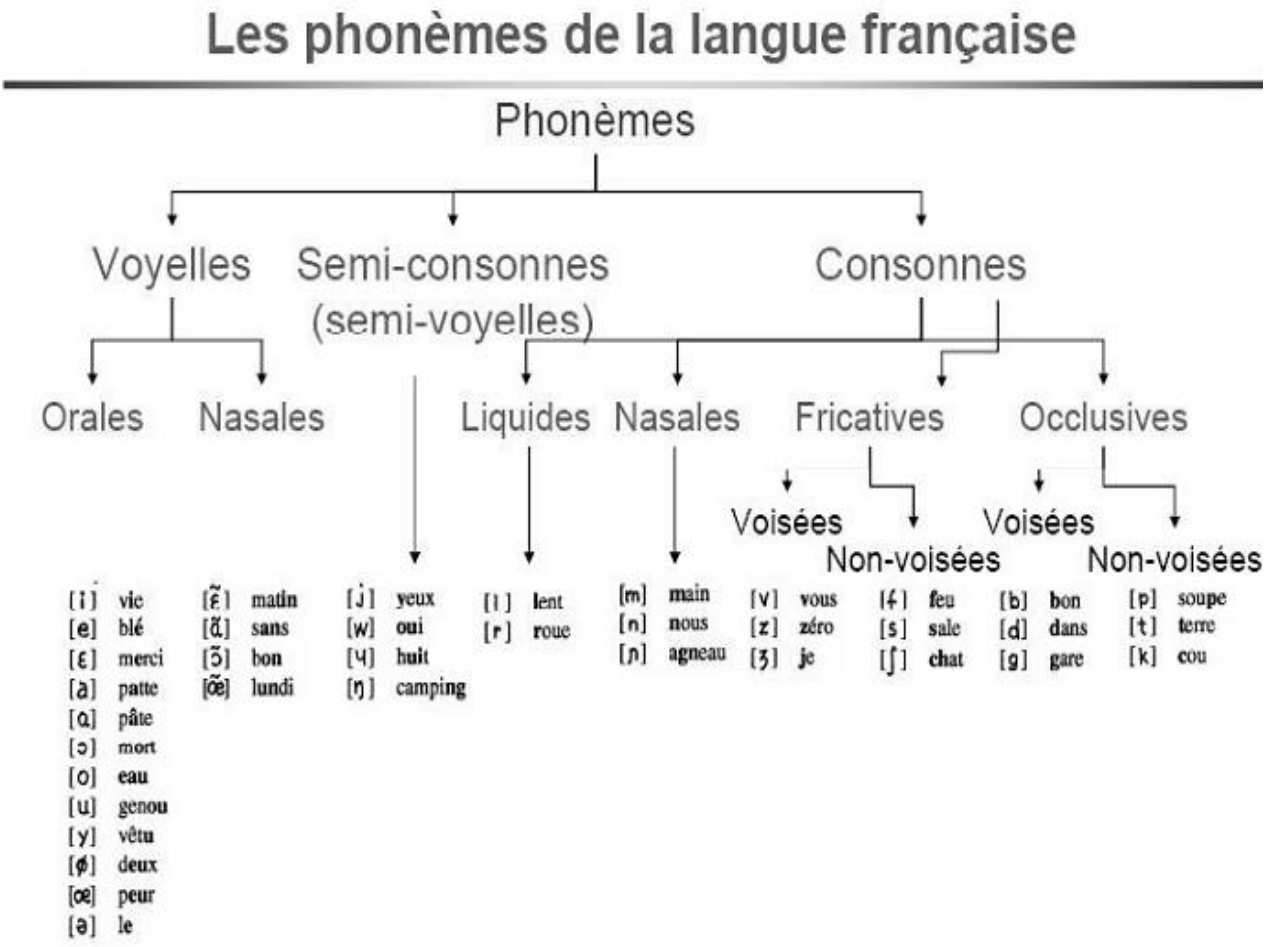


Figure I.11 Phonème de la langue française

I.3.1 Les voyelles

La caractéristique majeure des voyelles est le libre passage de l'air à partir des cavités supra glottiques.

Le seul traitement que l'air peut dès lors subir est la résonance (c'est-à-dire le renforcement de certaines bandes de fréquences).

Le timbre d'une voyelle dépendra de la variation des éléments suivants :

- ✓Le nombre des résonateurs mis en résonance (buccal, labial et nasal).
- ✓La forme du résonateur buccal.
- ✓Le volume du résonateur buccal.

Nombre de résonateurs :

On dénombre trois résonateurs : le résonateur buccal, le résonateur labial et le résonateur nasal. Le nombre de résonateur est toujours au moins de un, puisque le résonateur buccal est toujours présent.

Si le voile du palais est relevé, l'air ne traverse pas le résonateur nasal, mais se répand exclusivement dans le résonateur buccal, si le voile du palais est abaissé, l'air traverse simultanément les résonateurs buccal et nasal, ce qui porte à deux au moins le nombre de résonateurs.

Si les lèvres sont projetées vers l'avant et arrondies, il se forme un troisième résonateur à la sortie du canal buccal le résonateur labial ; si au contraire les lèvres sont appliquées contre ; les dents, le résonateur labial ne se forme pas.

D'après les critères ci-dessus, on oppose :

- ✓ des voyelles nasales (présence du résonateur nasal) à des voyelles orales (absence du résonateur nasal).
- ✓ des voyelles arrondies (présence du résonateur labial) à des voyelles non arrondies (absence du résonateur labial).

Classement des voyelles :

À l'intérieur de la catégorie des voyelles, le classement se fait selon :

- ✓ la nasalité (voyelles nasales ou orales)
- ✓ l'aperture ou ouverture du conduit vocal qui dépend de l'élévation de la langue par rapport à la voûte palatine (voyelles fermées ou ouvertes).
- ✓ la zone d'articulation déterminée par la position du dôme de la langue dans la cavité buccale (voyelles antérieures ou postérieures).
- ✓ la forme des lèvres (voyelles arrondies ou non arrondies).

L'utilisation systématique de ces critères permet de définir les voyelles cardinales (orales) dont le trapèze vocalique fournit une représentation schématique.

I.3.2 Les consonnes et les semi-voyelles

Les consonnes se différencient des voyelles par la présence d'un obstacle qui empêche le libre écoulement de l'air. La qualité de cet obstacle, ou mode d'articulation, est le critère principal qui permet de les distinguer entre elle. Le second critère de classification est la position de cet obstacle, ou point d'articulation.

Modes d'articulation :

En phonétique articulatoire, le mode d'articulation d'une consonne désigne un ensemble de propriétés de son articulation qui modifient la nature du courant d'air expiré. Il existe deux grands modes d'articulation consonantique :

- ✓ soit le passage de l'air est fermé (occlusion momentanée du chenal expiratoire) et le son résulte de son ouverture subite, on a alors affaire à des consonnes occlusives (cf. [k]) ; les consonnes occlusives sont des sons bruités de courte durée, caractérisés par un silence provenant de la fermeture complète du conduit vocal en un point précis.
- ✓ soit le passage se rétrécit mais n'est pas interrompu; on parle dans ce cas de consonnes continues ou constrictives dont les fricatives sont les plus représentatives (cf. [s]) ; les consonnes fricatives sont des sons bruités produits par l'écoulement turbulent de l'air : lorsque cet écoulement rencontre un rétrécissement, un lieu de constriction, il se produit un bruit de friction.

On distingue les consonnes orales des consonnes nasales, selon la cavité de résonance utilisée :

Au carrefour du pharynx, le passage de l'air peut en effet s'effectuer dans une ou deux directions, selon la position du voile du palais :

- ✓ si le voile du palais est relevé, l'accès aux fosses nasales est bloqué, et l'air ne peut traverser que la cavité buccale, la consonne est orale.

- ✓ si le voile du palais est abaissé, une partie de l'air traversera les fosses nasales (l'autre partie poursuivant son chemin à travers la cavité buccale), la consonne est: nasale (comme [n] dans « nous »).

Le français comporte les consonnes nasales suivantes :

-Bilabiale : [m] (maman)
-Dentale : [n] (ni)
-Palatale (en voie de disparition) : [ɲ] (gnangnan)
-Vélaire (mots d'emprunt) : [ŋ] (parking)

Consonnes occlusives :

Les consonnes occlusives (ou encore explosives) sont donc produites par une fermeture complète du chenal respiratoire, et non un simple rétrécissement, ce qui les différencie des continues.

L'occlusion se fait en deux temps :

- ✓ arrêt de la colonne d'air par la fermeture soudaine du chenal expiratoire.
- ✓ libération de l'air interne par le relâchement brusque de l'occlusion.

Consonnes fricatives :

Les consonnes fricatives (ou constrictives) sont donc produites par un resserrement du chenal expiratoire qui ne va pas contrairement à ce qui se passe pour les occlusives jusqu'à fermeture complète. Ce sont essentiellement les lèvres et la langue qui, selon leur position et leur tension musculaire particulière, conditionnent le type de friction réalisée.

Par exemple :

Le son [f], qui peut être soit écrit f ou bien ph dans les mots à racine grecque est une consonne fricative labiodentale sourde.

- ✓ le [v], qui peut être soit écrit v ou bien w dans les mots d'origine germanique est une fricative labiodentale voisée.

- ✓ le [s] est une consonne fricative alvéolaire sourde.
- ✓ le [z], qui s'écrit s entre voyelles ou z est une consonne fricative alvéolaire voisée.

I.3.3 Exemple

Représentation temporelle de «le chapeau» avec un léger bruit de fond.

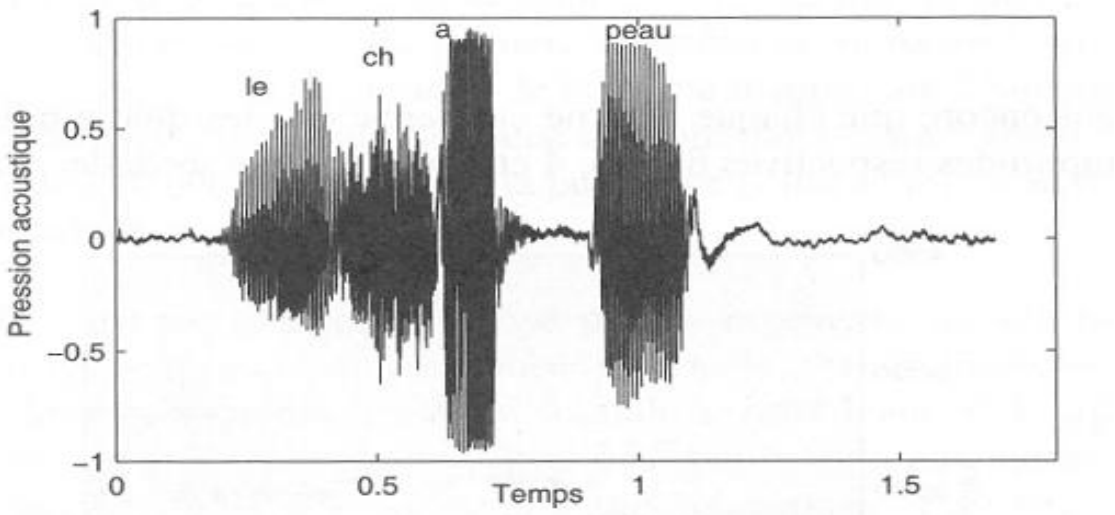


Figure I.12 Représentation temporelle du mot « le chapeau » [21]

1.4 Conclusion

Ce chapitre met en évidence la complexité du signal acoustique (parole) en illustrant toutes les caractéristiques physiques de ce dernier, et touts les organes qui participent dans la chaine de la production et la manière avec laquelle ils interviennent dans ce phénomène.

Une fois la parole est produite, un autre système prend le relais, il s'agit du système auditif qui va jouer le rôle d'un décodeur acoustique.

CHAPITRE II

Perception auditive et perception de la parole

Dans le cadre de notre recherche, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production.

L'oreille humaine est un système d'analyse du son étonnant et complexe.
Elle est capable de détecter des sons sur une large plage d'intensités et de fréquences.

II.1 Anatomie et fonctionnement

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pressions sont analysées dans *l'oreille interne* qui envoie au cerveau l'influx nerveux qui en résulte, le phénomène physique induit ainsi un phénomène psychique grâce a un mécanisme physiologique complexe. L'appareil auditif comprend, *l'oreille externe, l'oreille moyenne*, et *l'oreille interne*. Le conduit vocale relie le pavillon et le tympan : c'est un tube acoustique de section uniforme fermé a une extrémité. Son premier mode de

résonnance est situé vers 3000Hz, ce qui accroit la sensibilité du système auditif dans cette gamme de fréquences.

Le mécanisme de l'oreille interne (marteau, étrier, enclume) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la *cochlée* celle-ci contient la *membrane basilaire* qui transforme les vibrations mécaniques en vibrations nerveuses. La membrane s'élargit s'épaissit au fur et a mesure que l'on s'approche de l'apex de la cochlée, elle est le support de *l'organe de corti* qui est constitué par environ 25000 *cellules ciliées* raccordées au nerf auditif.

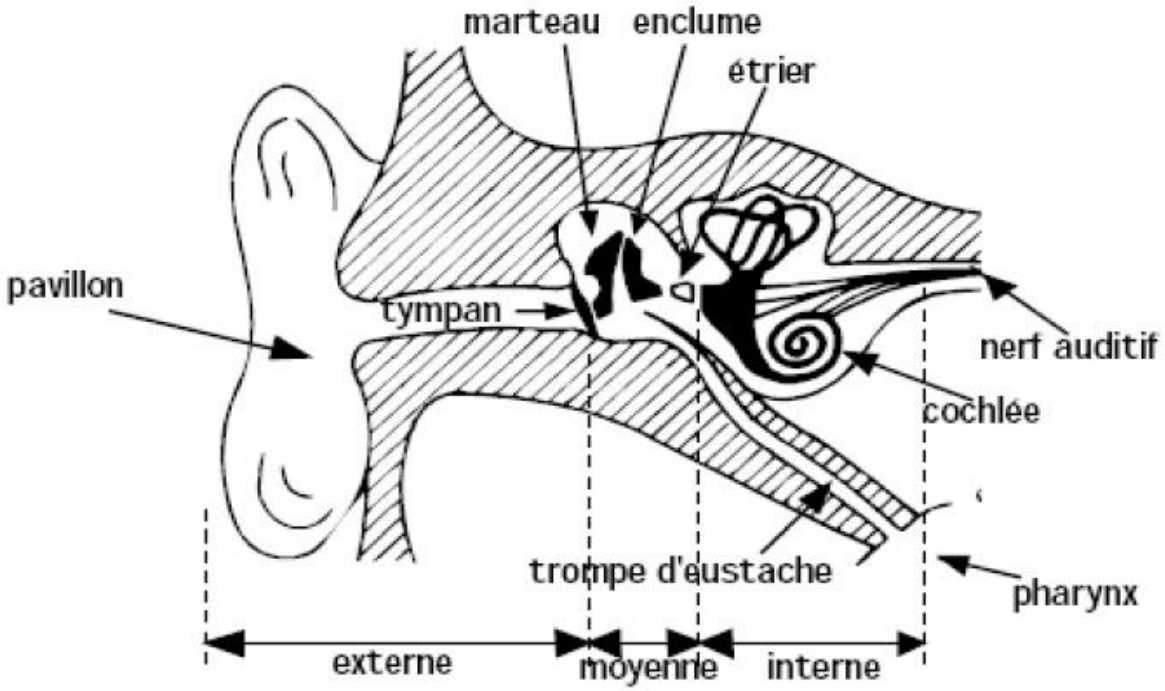


Figure II.1 Composition anatomique de l'oreille

II.2 Les courbes de réponses

La réponse en fréquence du conduit au droit de chaque cellule est représentée par la figure suivante.

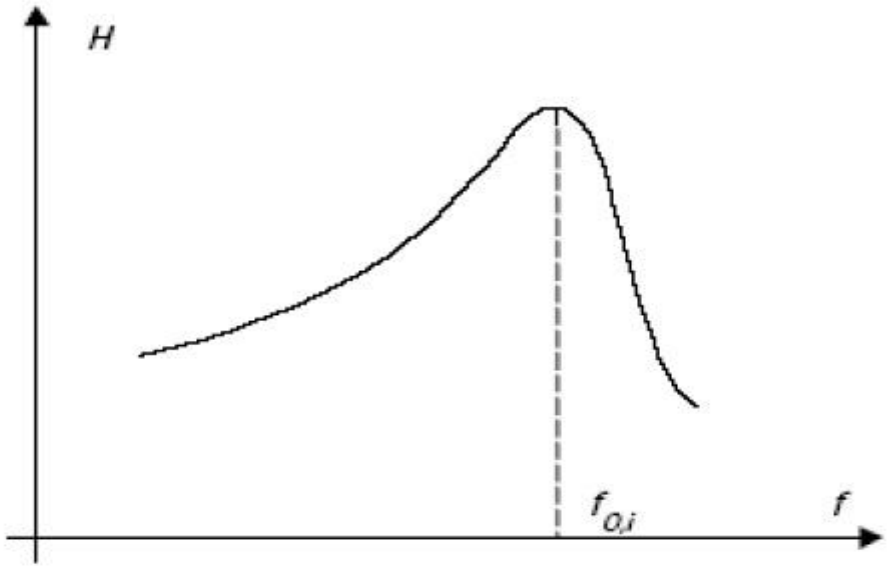


Figure II.2 Courbe de réponse du conduit vocale

La fréquence de résonance $f_{0,i}$ dépend de la position occupée par la cellule sur la membrane basilaire, au-delà de cette fréquence, la fonction de réponse s'attenue très vite.

Les fibres nerveuses aboutissent à une région de l'écorce cérébrale appelée *aire de projection auditive* et située dans le lobe temporal. En cas de lésion de cette aire, on peut observer des troubles auditifs. Les fibres nerveuses auditives afférentes (de l'oreille au cerveau) sont partiellement croisées, chaque moitié du cerveau est mise en relation avec les deux oreilles internes.

L'oreille ne répond pas également à toutes les fréquences, la figure suivante présente le champ auditif humain, délimité par la courbe de *seuil de l'audition* et celle *du seuil de la douleur* sa limite supérieure en fréquence (~16000Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (~32000Hz).

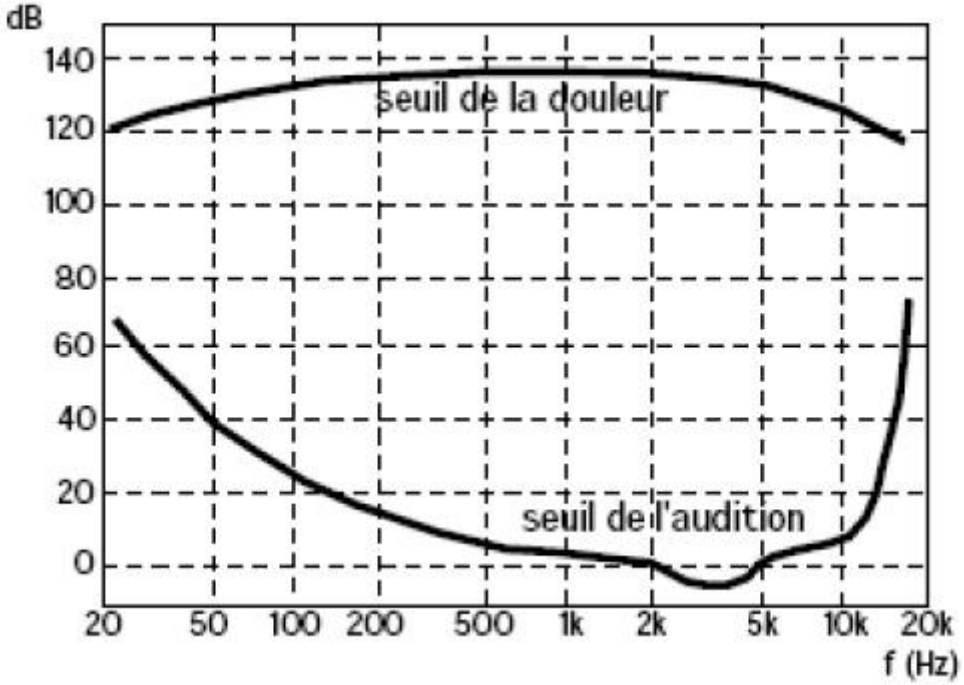


Figure II.2 Seuil de l'audition et de douleur

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences, la figure suivante fait apparaitre les courbes d'égales impressions de puissance auditive-physiologie auditive (aussi appelée *sonie*, exprimée en *sones*) en fonction de la fréquence. Elles relèvent un maximum de sensibilité dans la plage [500Hz-10kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

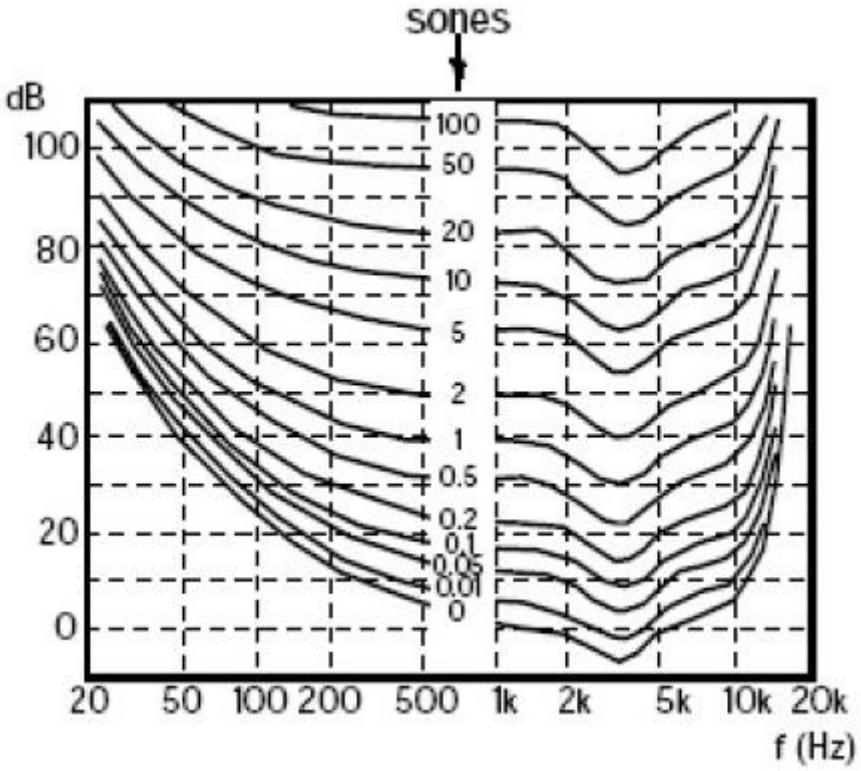


Figure II.3 Sensibilité de l'oreille

Enfin, un son peut cacher un autre. Cette priorité psycho acoustique, appelée *phénomène de masquage*, peut être visualisée sous la forme de courbes de masquage (Figure II.3) qui mettent en évidence la modification locale du seuil de l'audition en fonction de la présence d'un signal déterminé.

II.3 Conclusion

Remarquons pour terminer que ce qui est perçu n'est pas forcément *compris*, une connaissance de la langue interfère naturellement avec les propriétés psycho acoustique de l'oreille. En effet, les sons ne sont jamais prononcés isolément, et le contexte phonétique dans lequel apparaissent est lui aussi mis à contribution par le cerveau pour la compréhension du message. Ainsi, certains sons portent plus d'information que d'autres, dans la mesure où leur probabilité d'apparition à un endroit donné de la chaine parlée est plus faible, de sorte qu'ils réduisent l'espace de recherche pour les sons voisins.

Les sons sont organisés en unités plus larges, comme les mots, qui obéissent eux-mêmes à une syntaxe et constituent une phrase porteuse de sens.

CHAPITRE III

Accès au lexique et syntaxe

L'une des caractéristiques essentielles des processus de compréhension du langage est la rapidité de leur mise en œuvre. Les énonces sont compris au fur et a mesure de leur réception, c'est-à-dire en « temps réel ». Compte tenu de ce qui précède, il semble donc nécessaire d'accepter qu'une partie importante du travail de compréhension prenne place lors de la lecture ou de l'audition des mots qui composent l'énoncé. L'étude des mécanismes psychologiques impliqués dans l'identification des mots est l'un des objectifs majeurs de la psycholinguistique.

Bien que l'étude de l'identification des mots ait depuis longtemps constitué l'un des domaines privilégiés de la psychologie expérimentale, on assiste actuellement à un renouvellement et à un développement extraordinaire des travaux consacrés à ce thème. Ce renouvellement est lié de manière étroite à l'insertion de ces études dans le champ de la psycholinguistique par le truchement de la métaphore du lexique mental ou lexique interne. [2]

Accepter l'hypothèse d'un lexique subjectif en tant que système dans lequel sont représentées toutes les informations concernant les mots de la langue conduit à soulever plusieurs questions importante dont les suivantes :

- ✓ Quelles sont la nature et l'organisation de ce lexique ?
- ✓ Quel est le format des représentations lexicales ?
- ✓ Quelles sortes de procédures rendent possible l'accès à ces représentations ?

Ces différentes questions sont sans doute fortement reliées car il est clair qu'il va s'avérer difficile de préciser les procédures d'accès sans formuler les hypothèses concernant les principes généraux d'organisation du lexique. Quoi qu'il en soit, nous examinerons dans ce chapitre la problématique de l'accès au lexique. [2]

Pour aborder l'étude de l'accès, il est utile de distinguer les procédures d'accès au lexique de celles d'utilisation des informations continues dans ce lexique. Dans le domaine de la perception du langage, l'accès peut être envisagé comme le résultat des opérations qui permettent d'associer une représentation sensorielle à une représentation mentale correspondant a un mot de la langue. [2]

III.1 La syntaxe

La syntaxe est l'étude de l'organisation des termes de la phrase. La phrase ayant une structure discrète, les questions qui se posent sont de nature combinatoire, du point de vue théorique aussi bien que du point de vue expérimental.

III.1.1 La grammaire générative de CHOMSKY

En 1957 Chomsky propose une nouvelle théorie, beaucoup plus computationnelle que les grammaires classiques, fondée sur la décomposition structurelle en constituants (phrase, groupes, syntagmes, mots, etc.) formalisable par des règles de réécriture. Partant de l'hypothèse que la syntaxe est « innée », il suggère qu'il existe une structure de surface (la

phrase) et une structure profonde déduite de cette dernière par des règles de transformation. Ses travaux suscitent beaucoup de recherches en psycholinguistique mais nul ne parvient à fonder sa théorie chez l'humain. Il la révise alors complètement (1982) en introduisant le « liage » directement sur la structure de surface. Cette théorie parait être ad hoc et s'éloigne encore plus d'un modèle cognitif possible. Pour rendre l'étude de la syntaxe indépendante du sujet parlant, il avait dès le début distingué performance et compétence : la compétence serait celle d'un locuteur parlant parfaitement la langue en connaissant toutes les règles, la performance serait le niveau de réalisation réellement atteint par un sujet parlant.

III.1.2 Les grammaires lexicales fonctionnelles de BRESNAN-KAPLAN

La différence avec la grammaire de Chomsky (première manière) réside essentiellement dans l'absence de distinction entre structure profonde et structure de surface. La syntaxe permet d'attribuer à toute phrase de la langue:

(a) une structure de constituants (c-structure) qui est un arbre étiqueté engendré a partir de règles de réécriture mais qui décrit directement l'agencement superficiel des éléments de la phrase.

(b) une structure fonctionnelle (f-structure) qui sera le seul input de la composante sémantique et qui distribue les notions "SUJET", "OBJET", "ADJOINT", "GENRE", "NOMBRE", etc.

Il n'y a pas de règles transformationnelles. Le lexique contient des informations catégorielles et fonctionnelles. Les règles de réécriture peuvent être annotées à l'aide de "gabarits" qui permettent de colporter des informations fonctionnelles dans l'arbre syntaxique.

Ex: - la fille prend au bébé le jouet
- la fille prend le jouet au bébé

Le verbe prendre accepte les deux constructions:

-CONST prendre <- (+SUJ) (+A-OBJ) (+OBJ)>

- CONST prendre < -(+SUJ) (+OBJ) (+A-OBJ)>

et une règle unifie les deux solutions

(+OBJ) <-> (+A-OBJ) (ce n'est pas une règle transformationnelle)

L'équivalence sémantique se retrouve au niveau de l'interprétation sur les notions de "agent", "bénéficiaire", "objet", etc. et non après un jeu subtil de transformations syntaxiques comme chez Chomsky. Les règles de constituants sont pour cet exemple:

P GN. GV   /(+SUJ)=-, +=-/

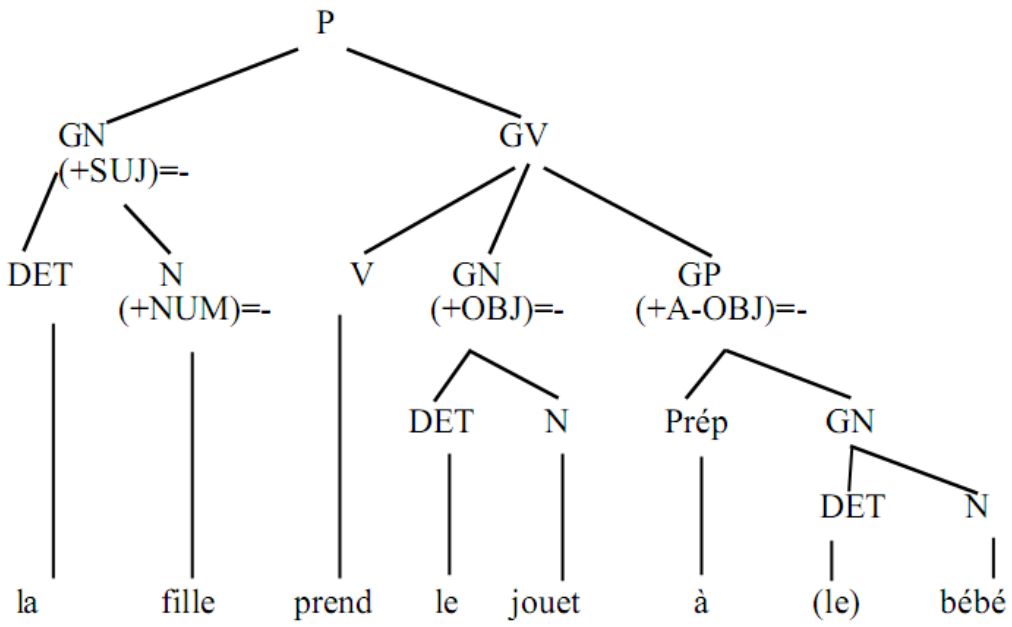
GN -> DET. N / ( ), (+NUM)=-/

GV -> V. GN .GP   / ( ), (+OBJ) = -, (+A-OBJ) = - /

GP -> Prep. GN

+ et - notent les relations de dominance sur le père (+) et les fils (-) qui sont ‡ instancier au niveau de la structure fonctionnelle. SUJ, OBJ etc. sont des meta variables.

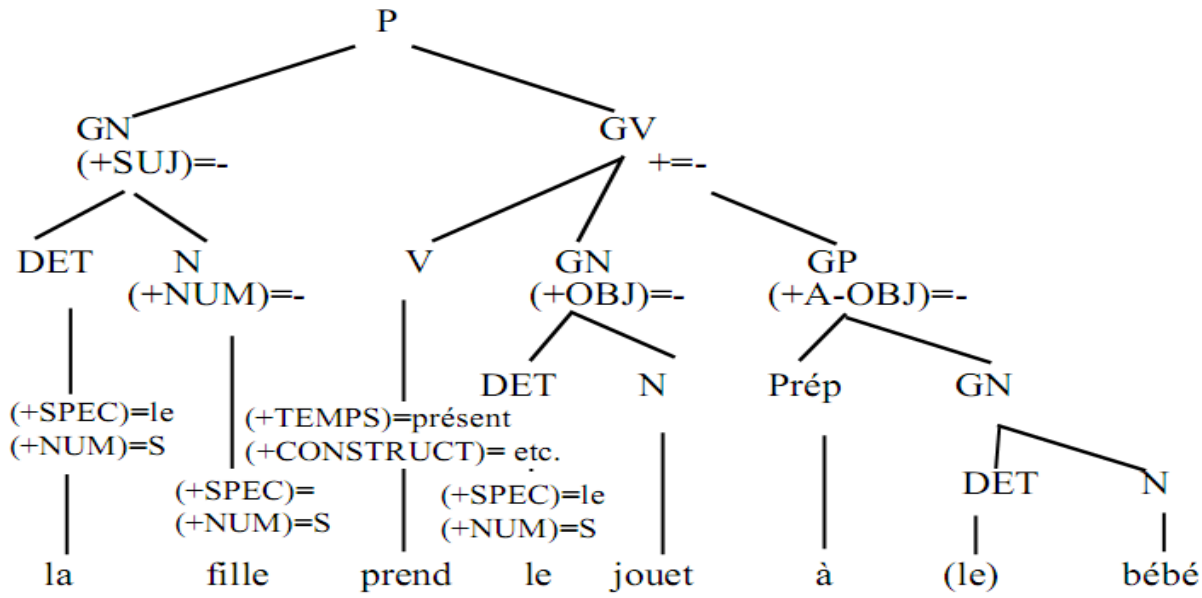
Représentation de l'arbre syntaxique "décoré":



Représentation du lexique :

```
la      DET, (+SPEC)=le, (+GENRE)=F, (+NUM)=S
fille   N, (+SEM)="Personne", (+GENRE)=F, (+NUM)=S
prendre V, (+TEMPS)="Présent", (+MODE)="Ind", (+PERS)=3, (+NUM)=S,
        CONSTRUCT prendre <(+SUJ) (+A-OBJ) (+OBJ)>
le      DET, (+SPEC)=le, (+GENRE)=M, (+NUM)=S
jouet   N, (+SEM)="Objet", (+GENRE)=M, (+NUM)=S
au      DET, (+SPEC)=à le, (+GENRE)=M, (+NUM)=S
bébé    N, (+SEM)="Personne", (+GENRE)=M, (+NUM)=S
```

L'analyse se fait en propageant les valeurs des méta-variables jusqu'au lexique (ascendant ou descendant selon les indications fournies par les opérateurs + et -), puis en remontant le long des nœuds, on s'assure de la cohérence des attributs.



Ce qui donne l'analyse suivante:

SUJ=fille

SPEC=le

NUM=singulier

GENRE=féminin

SEM=personne

ACTION=prendre

TEMPS=présent

MODE=indicatif
PERS=3
NUM=singulier
COSTRUCTION prendre <(+SUJET) (+A-OBJ) (+OBJ)>

OBJ=jouet
SPEC=le
NUM=singulier
GENRE=masculin
SEM=objet

A-OBJ=bébé
SPEC=à le
NUM=singulier
GENRE=masculin
SEM=personne

III.2 Le lexique

III.2.1 Les catégories lexicales

On distingue deux grandes classes lexicales

Mots outils (ou mots grammaticaux) :
- Déterminants
- Pronoms
- Prépositions
- Conjonctions
- Interjections

Mots lexicaux :

Adverbes

Noms

Adjectifs

Verbes

Locutions particulières (lexicalisées)

La première catégorie est peu évolutive et constante dans presque tous les champs d'application. La seconde est très variable en ce qui concerne la quantité des instances des classes : c'est la richesse du vocabulaire.

III.2.2 Pronoms et déterminants

a) **Pour les déterminants** :

Du premier groupe D1

Articles

Défini	le, la, les, [à+] au, aux, du, des
Indéfini	un, une, de, des
Loc indéf	. N'importe lequel, je ne sais quel
Partitif	du [de le], de, la, des
Démonstratifs	ce, cet, cette, ces, ce+ci[là]
Possessifs	mon, ton, son, notre, votre, leur, ma, ta, sa, nos, vos, leurs
Interrogatifs, Exclamatifs	quel, quelle, quels, quelles
Relatifs	lequel, laquelle, duquel, de laquelle, lesquels, lesquelles

Du deuxième groupe D2

Numéraux (quantitatifs)

Cardinaux un, deux, trois...etc.

Ordinaux premier, deuxième...etc.

Formes substantiv2es

Multiplicatifs double, triple, centuple...

Fractions moitié, tiers, quart...

Collectifs dizaine, douzaine, centaine...

Indéfinis

Quantitatifs

Nulle nul, aucun, pas un

Singulier tout, chaque, tel, quelque, un certain, maint

Pluriel quelques, certains, maints, plusieurs, divers, différents

Indéterminée quelque

Totalité numérale tous

Totalité quantitative tout (est aussi adverbe)

Loc.la plupart, une foule, une multitude, une masse de

Identité

Différence

b) Pour les pronoms :

Personnels

Formes normales je, tu, il, nous, vous, ils, elle, eux, elles, lui, on

Formes réfléchies

Antéposées me, te, se, s', nous, vous

Postposées moi, soi, toi, lui, nous, vous, eux

Démonstratifs c', ce, celui, cela, ce+ [ci, là], ce [lui, eux, elle, elles] + [ci ,là]

Présentatif c'est...

Relatifs

Simple qui, que, quoi, qu', dont, où

Composé [le+, à+, de+] quel, qui que ce soit, qui, quiconque

Interrogatifs

Formes simples qui, que, quoi

Formes composées [à+, de+, par+, pour+, sur+] qui, quoi, qui est-ce qui ...

Indéfinis qui, quiconque (qui vivra verra), quel que, quel que soit, celui que, quoi que ce soit qui, quoi que....

Possessifs le mien, le tien, le sien, le nôtre, le vôtre, le leur

III.2.3 Préposition, conjonction et interjection

a) Pour les prépositions

Racines

1er groupe de, à

2ème groupe dans, par, pour, sur, avec, devant, derrière, sans, sous, contre, vers, chez

Locutions prépos. Loin de, près de, au lieu de, à côté de [='de' + A ou GP]

b) Pour les conjonctions

Coordination et, ou, ni, mais, or, car, donc

Subordination

Complétives que

Circonstanciels si, quand, comme, lorsque, puisque, quoique, parce, que, dès que, pour que, etc.

Locutions conj. (construites avec 'que')

Temps aussitôt que, avant que, etc.

But pour que, afin que, etc.

Cause parce que, du fait que, etc.

Conséquence de sorte que, pour que, etc.

Concession/opposition bien que, alors que, etc.

Condition pourvu que, à condition que, etc.

Comparaison ainsi que, de même que, etc.

c) Pour les interjections

Hé, quoi, etc.

III.2.4 Verbe, nom et adverbe

a) Pour les verbes

Catégories syntaxiques

Transitif

Direct

Indirect

Avec 2 C.O.D. possibles (+combinatoire animé/non-animé)

Intransitif

Copules: être, devenir, rester, sembler, paraître, etc.

Conjugués avec être

Conjugués avec avoir (et/ou être)

Pronominaux se regarder, etc.

Auxiliaires: devoir, falloir, être, avoir, pouvoir, aller, venir, etc.

Sujet animé: Penser, marcher, etc.

Catégories sémantiques

Procès (action)

Factitif (peut être remplacé par 'faire'+infinitif)

Mouvement du robot: avancer, reculer, tourner, attendre, arrêter, contourner, longer, monter, suivre, descendre, aller, passer, accélérer, ralentir, etc.

Manipulation: prendre, lâcher, mesurer, déplacer, bouger, lire, écrire, allumer, éteindre, détruire, etc.

Perfectif/imperfectif (passé encore actuel ou non) comprendre, finir, partir, arriver, etc. /posséder, espérer, hésiter, habiter, etc.

Duratif/momentané itérer, recommencer, continuer, donner, rester, /copier, transmettre, alerter, appeler, sauter, etc.

b) Pour les noms

Peuvent être classés dans une matrice de traits:

commun/propre, animé/non-animé, humain/non-humain, concret/abstrait, comptable/non-comptable, simple/composé, sur composé, individuel/collectif

Espèce

Robot, roue, capteur, bras, etc.

Environnement fixe obstacle, bâtiment, rue, lampadaire, etc.

Environnement mobile table, chaise, lampe, tuyau, etc.

Lieu

De passage: porte, couloir, escalier, montée, etc.

Obstacles: mur, sol, coin, etc.

Temps: jour, heure, minute, seconde, date, instant, etc.

Instrument: clé, bras, capteur, etc.

Matière: bois, fer, ciment, etc.

Indéfinis

Unités de mesure: centimètre, mètre, kilo, etc.

Directions: gauche, droite, haut, bas, est, ouest, etc.

Qualité (comportement d'adjectif)

D'action: avancée, recul, etc. (nominalisation des verbes)

Etc.

c) Pour les adverbes

Manière Adj + [ment]

Lieu ici, devant, derrière

Temps hier, demain

Quantité

Modificateurs

Comparatifs plus, moins, autant, aussi

Superlatifs très, trop, peu, beaucoup, très bien, passable, médiocre

Interrogatifs, exclamatifs comment, où, quand

Négation pas, point, guère, plus, jamais, personne, rien, aucun, nul

Opinion oui, si, non

Modalisateurs

Insistance certainement

Réserve probablement, peut-être

Liaison ensuite, puis, ainsi, en effet, aussi

Locutions adverbiales de coordination à l'improviste, à tort et à travers, à l'anglaise, à croupetons, à califourchon...

Interrogatifs

Lieu	où
Cause	pourquoi
Temps	quand
Manière	comment
Quantité	combien

III.2.5 Adjectifs et participes

Peuvent être classés dans une matrice de traits:
commun/propre, animé/non-animé, humain/non-humain, concret/abstrait, comptable/non-comptable, simple/composé, etc. comme les noms

Qualification

Manière	incurvé, souple, silencieux, etc.
Temps	tard, tôt, etc.
Lieu	éloigné, rapproché, etc.
Matière	ferreux, poussiéreux, etc.
Forme:	allongé, écourté, etc.
Couleur	
Simple	vert, rouge, etc.
Composée	rouge foncé, bleu clair, etc.
Dimension	grand, petit, étroit, etc.
Poids	lourd, léger, etc.

Concret/abstrait	
Animé/non animé	
Etat permanent/passager	haut, bas/sale, propre
Transitif/intransitif	

III.3 Le code phonétique

La représentation phonétique d'une unité acoustique est très importante, elle nous permet de voir la structure interne du mot et savoir la représentation orthographique de cette

unité. Ca nous aidera par la suite a élaboré un modèle linguistique basé sur cette représentation des mots.

On utilise 34 phonèmes pour l'écriture phonétique, y compris le silence. Le tableau suivant donne la liste des phonèmes avec un mot clé et une codification utilisée pour chacun.

PHONEME	MOT CLE	CODE
Silence		1
a	Plat	2
r	rue	3
l	lent	4
é	blé	5
s	sous	6
i	il	7
è	lait	8
e	le	9
k	cou	10
t	ta	11
p	pas	12
d	dans	13
m	ma	14
$\hat{a}$	au	15
n	nous	16
u	ou	17
v	vie	18
y	nu	19
$\hat{o}$	on	20
o	eau	21
z	je	22
$\hat{u}$	bol	23
$\hat{e}$	Lin	24
f	feu	25
b	bon	26

w	voir	27
î	huit	28
ĕ	Heure	29
Z	Zéro	30
J	Bailler	31
c	Chat	32
g	gare	33
ũ	un	34

Tableau III.1 Liste des phonèmes avec leur mot clé et la codification

III.4 Conclusion

Dans ce chapitre, nous avons vu les composantes lexicales et syntaxique, qui vont servir de module de vérification pour le système de reconnaissance et compréhension de la parole (module linguistique), elles lui permettront de vérifier la validité des phrases (ou des mots) qui seront reconnus par un autre module qui est le module acoustique.

Ce module acoustique a pour tache d'extraire les informations pertinentes (acoustiques) du signal audio (parole).

CHAPITRE IV

Outils pour le traitement de la parole

Un son acoustique est caractérisé par son amplitude, sa durée, et de son timbre. La parole est un son particulièrement complexe. Le traitement du signal a pour but précisément de quantifier ces trois grandeurs pour faire correspondre a l'onde sonore (temporelle) une description multidimensionnelle. [2]

Pour quoi analyser la parole ?

Pour étudier et comprendre les phénomènes physiques mis en jeu

– être un peu moins ignorants...

– comprendre aussi les dysfonctionnements (handicapés du langage)

– être capable d'utiliser ces connaissances pour l'apprentissage des langues étrangères Pour reproduire la parole sous forme artificielle

– synthèse de la parole (synthèse à formants)

– modélisation de l'appareil de production Pour déterminer des mesures de caractérisation pouvant être utilisées pour le codage ou par les moteurs de reconnaissance de parole.

– caractéristiques spectrales (LPC, MFCC, fréquence fondamentale, etc.…) [2]

IV.1 Description du signal vocal

IV.1.1 L'audiogramme

La parole est un signal réel, continu, d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps : tantôt périodique (pseudopériodique) pour les sons voisés, tantôt aléatoire pour les sons fricatifs, tantôt impulsionnelle dans les phrases explosives des sons occlusifs. Cette structure reflète l'organisation temporelle des gestes de production et sur l'onde sonore apparaissent quelques caractéristiques de la source et du conduit vocal

- Pour la source : période fondamentale $T_0 = 2/F_0$ et amplitude A_0
- Pour le conduit : période des formants $T_i = 1/F_i$ i= 1,2 … ; amplitude A_i

Exemples :

a)

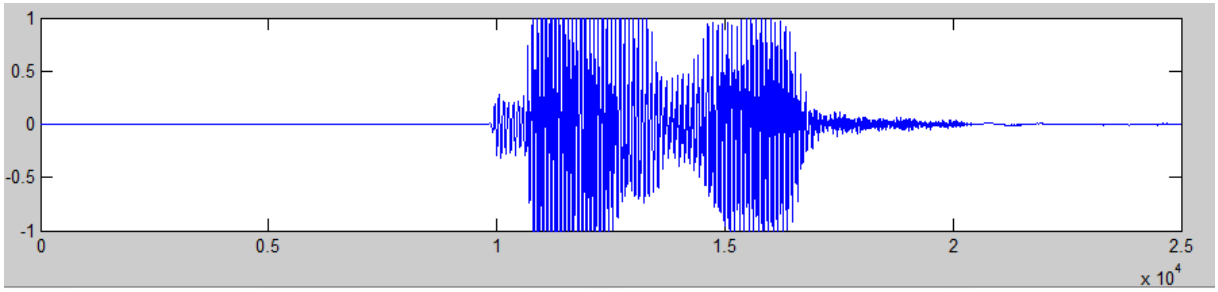


Figure IV.1 Audiogramme du mot « BONJOUR »

b)

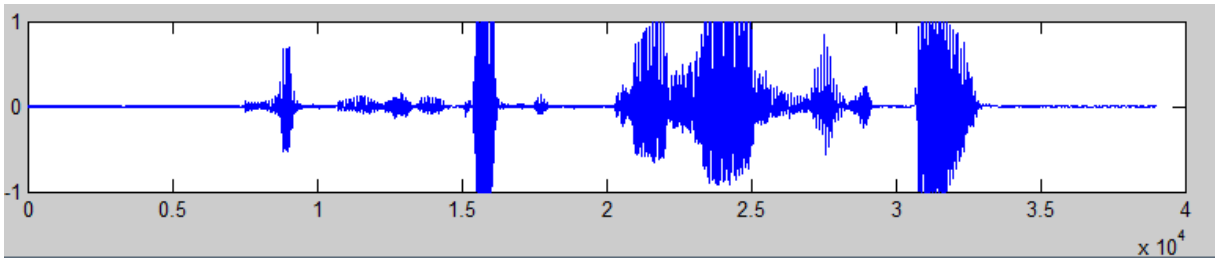


Figure IV.2 Audiogramme de la phrase «Vous êtes Monsieur KACETE amine n'est-ce pas ? »

IV.1.2 La variabilité

Une caractéristique importante du signal parole est la variabilité :

- une personne ne prononce jamais deux fois le même son de la même façon
- deux personnes ne prononcent pas le même son de la même façon

Pourtant, ce son est parfaitement reconnu et compris par un auditeur humain.

Exemple 1 :

Même sons, même personne, mêmes conditions d'enregistrement et prononciation du même mot :

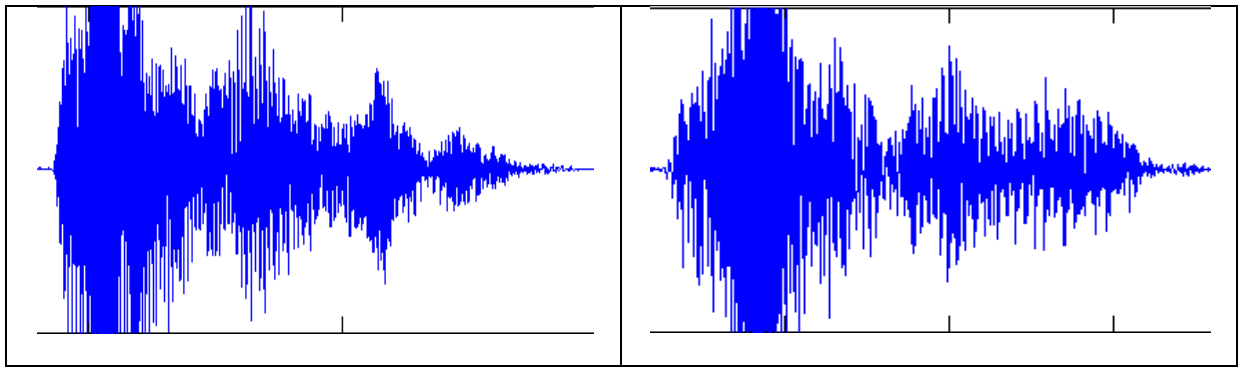


Figure IV.3 Variabilité intra-locuteur

Exemple 2 :

Même son, mêmes conditions d'enregistrement deux locuteurs différents et prononciation du même mot :

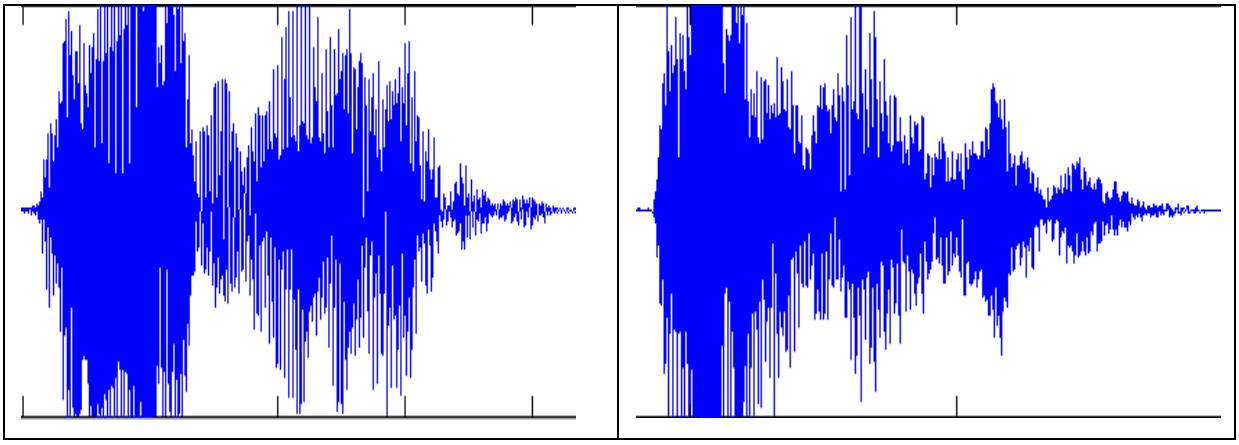


Figure IV.4 Variabilité interlocuteur

IV.2 Échantillonnage du signal vocal

Avec le développement des calculateurs (et des circuits numériques spécialisés) le traitement analogique du signal a subit un déclin important vis-à-vis du traitement numérique. La stabilité et la précision des systèmes numériques n'est plus à démontrer, la seule limitation peut être dans certains cas, leur lenteur de calcul. Aussi, avant le traitement, est-il nécessaire de numériser le signal continu sortant du microphone ou d'un appareil d'enregistrement. Cette opération s'appelle échantillonnage du signal, l'opération inverse étant l'interpolation.

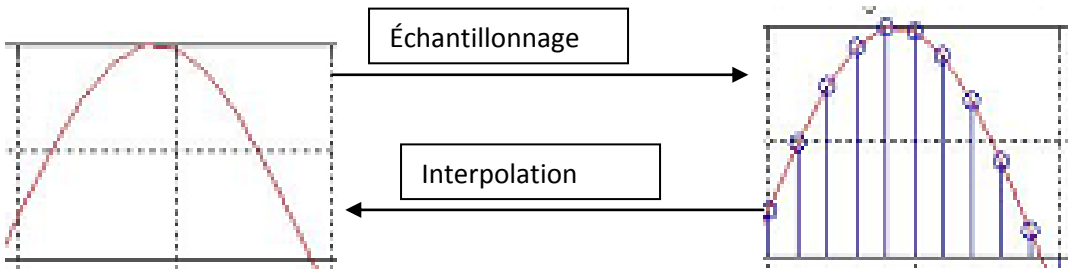


Figure IV.5 Echantilonnage et interpolation d'un signal.

IV.2.1 Définition

L'échantillonnage d'un signal analogique représenté par une fonction f(t) consiste à construire, à partir de f(t), un signal à temps discret f(n)= f ($n*t_e$) obtenu en mesurant la valeur de f(t) toutes les T_e secondes.

$$F(n) = f\,(n*T_e)$$

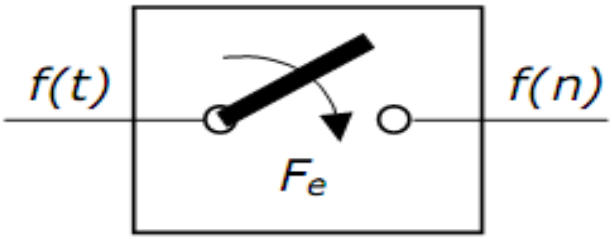


Figure IV.6 Représentation schématique de l'échantillonnage

Si f(t) subit une discontinuité par saut à un instant d'échantillonnage, on convient de poser :

$$f(n) = \frac{1}{2}[f(nT^{+}_{e}) + f(nT^{-}_{e})]$$

Le schéma de principe de l'échantillonnage est décrit à la Figure suivante. Il exprime le fait qu'on peut considérer que $f^{+}(t)$ est obtenu par multiplication de f(t) par un train d'impulsions de Dirac de période T_e :

$$f^{+}(t) = f(t)\delta(t) = \sum_n f(nT)\delta(t - nT)$$

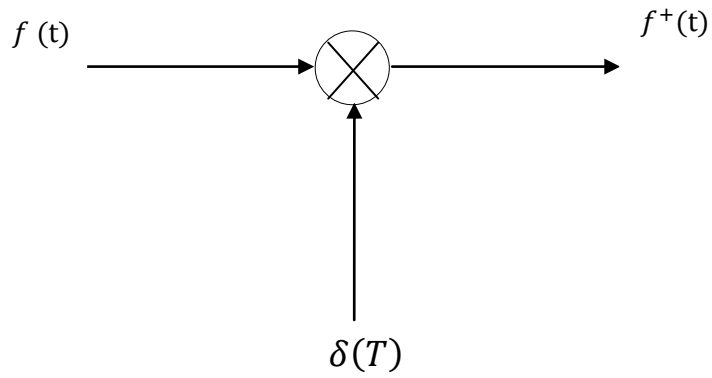


Figure IV.7 Représentation mathématique de l'échantillonnage

IV.2.2 L'échantillonnage dans le domaine fréquentiel

La transformée de Fourier :

-Instrument de base de la théorie du signal

- Représentation spectrale des signaux

- Exprime la répartition en fréquence de l'amplitude et de la phase de l 'énergie 'un signal

TF d'un signal continu

Soit x(t) un signal complexe, la TF est une fonction complexe de la variable réelle (ω=2πf) définie par :

$$F\{x(t)\} = X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-jax}dt$$

La transformée inverse est donnée par :

$$x(t) = F^{-1}\{X(\omega)\} = \int_{-\infty}^{+\infty} X(\omega) e^{+jax} d\omega$$

Exemple :

La TF du sinus

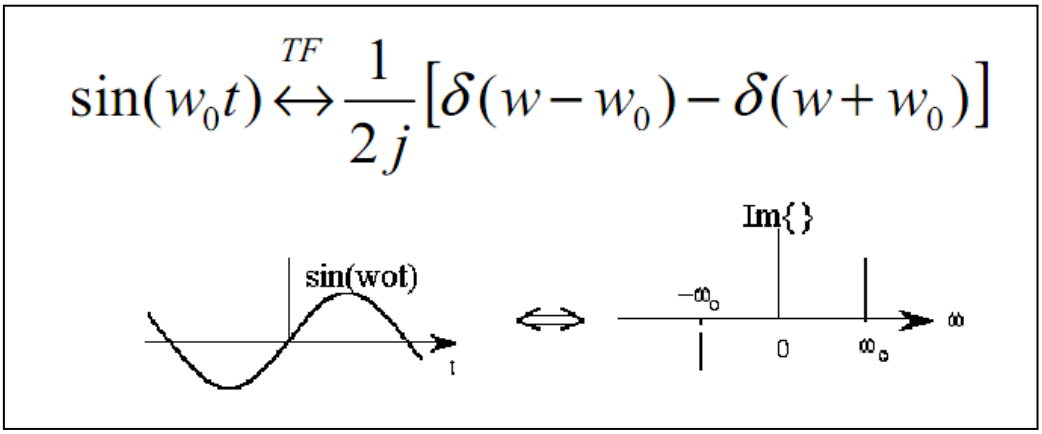


Figure IV.8 la TF de la fonction sinus

La transformée d'un sinus de fréquence ω_0 est une somme de 2 impulsions dans la partie imaginaire.

Spectrogramme temps-fréquence

$$S_x(t,f) = [\int_{-\infty}^{+\infty} x(s) h^*(s-t) e^{-2i\pi fs} ds]^2$$

Représentation du module du spectre du signal (donc de l'énergie) sur une fenêtre glissante temporelle h.

Exemple :

Cas d'une sinusoïde

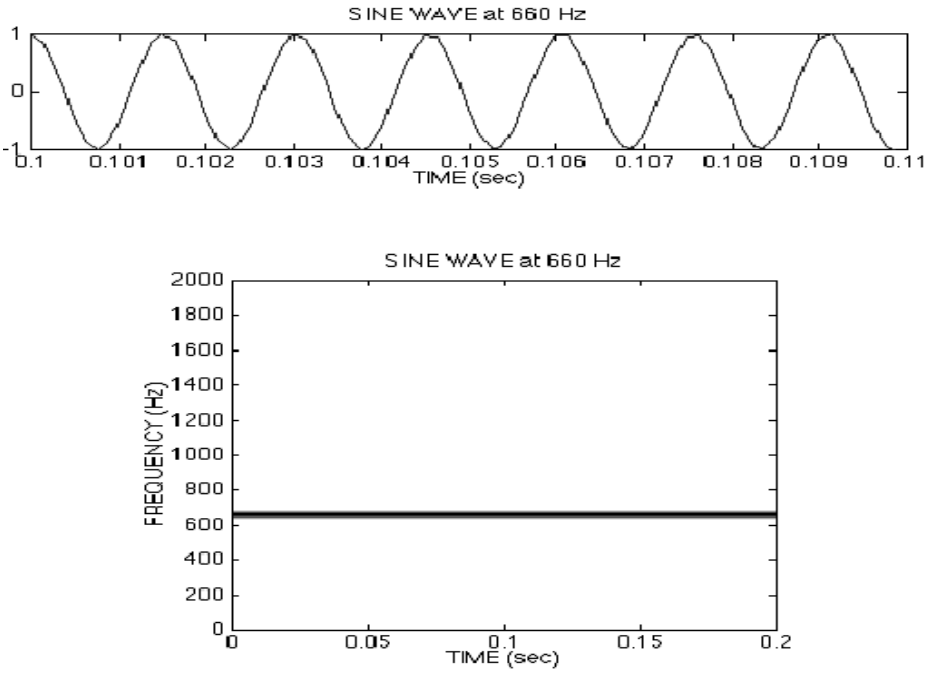


Figure IV.9 Représentation du module du spectre du signal sur une fenêtre glissante temporelle.

IV.2.3 Théorème de SHANNON

Un signal analogique x(t) ayant une largeur de bande finie limitée à 2F hz, ne peut être reconstitué exactement à partir de ses échantillons x(nΔt), que si ceux-ci ont été prélevés avec une période :

$$T_e = \frac{1}{f_e} \leq \frac{1}{2F}$$

Pour que la répétition périodique du spectre ne modifie pas le motif répété, il faut et suffit que la fréquence d'échantillonnage soit supérieure ou égale à 2 fois la fréquence maximum F du signal.

$$F \leq f_N = \frac{f_e}{2}$$

Exemples :

a) Signaux quantifiés sur 8 bits (256 valeurs possibles) et échantillonnés à 8kHz

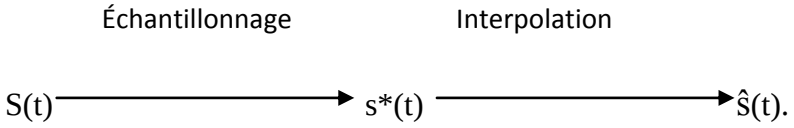
⇨ débit binaire = 64kbit/s

b) Signaux quantifiés sur 16 bits (65536 valeurs possibles) et échantillonnés à 16kHz

⇨ débit binaire = 256kbit/s

IV.2.4 L'interpolation dans le domaine temporel

C'est l'opération inverse de l'échantillonnage, a partir d'un signal s*(t) discret, lui faire correspondre un signal ŝ(t) continue en minimisant les pertes d'information au sens de Shannon.



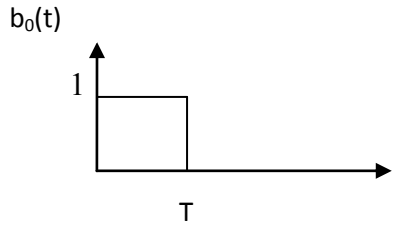
a) Interpolateur idéal : lorsque $f_m \leq f_e/2$ il est théoriquement possible de retrouver ŝ(t)=s(t) (théorème de Shannon). Pour cela considérons un filtre idéal passe-bas de fréquence de coupure $f_e/2$. Sa réponse impulsionnelle est p(t) = sin (π*t/T)/ (π*t/T) et le signal $\hat{s}(t)$ sortant de ce filtre, en ayant comme entrée s*(t), est :

$$\hat{s}(t) = \sum s_n \sin((\pi * t/T) - n * \pi)) / ((\pi * t) - n * \pi).$$

b) Interpolateur réalisable : en pratique puisque on s'éloigne de l'interpolateur idéal, on peut affirmer qu'il y'aura toujours une perte d'information lorsqu'on opère les transformations : s(t) ⟶ s*(t) ⟶ $\hat{s}$ (t).

Les interpolateurs les plus utilisés sont les bloqueurs d'ordre n (s(t) est interpolé a l'aide de n valeurs successives de s*(t)) et plus particulièrement du bloqueur d'ordre 0, c'est-à-dire :

$$\begin{cases} b_0(t) = 1 & 0 \le t \le T \\ b_0(t) = 0 & \text{ailleurs} \end{cases}$$



IV.3 Analyse spectrale de la parole

IV.3.1 Le spectrographe

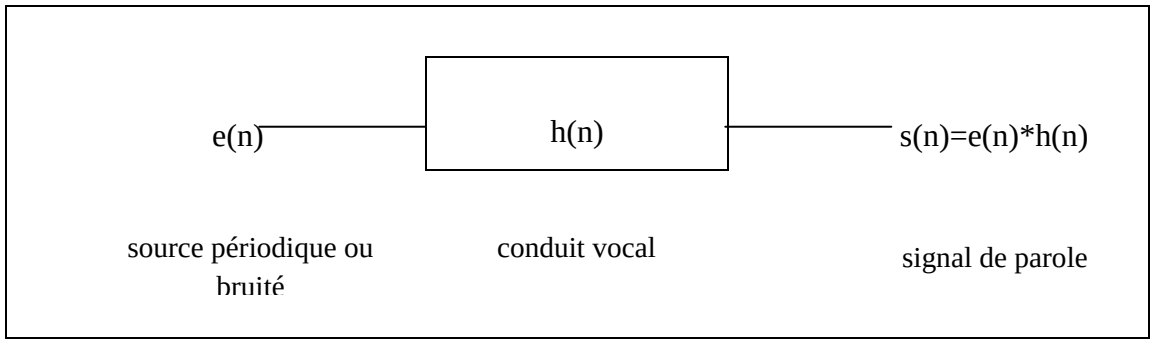
a) La transformée de Fourier discrète (TFD)

Considérons une suite finie de N échantillons {x(n)}={x(0), x(1),…, x(N-1)}. On définit sa transformée de Fourier Discrète comme la suite {X(k)} :

$$X(k) = \sum_{n=0}^{N-1} x(n)\, W_n^{-nk} \qquad (k = 0 \dots N-1)$$

$$avec\ W_N = e^{j\frac{2\pi}{N}}$$

b) Modèle simplifié de la production de la parole



Le spectre utile s'étend de 0 à 8Khz.

1. Caractéristiques de la source :

   – bruit pour les fricatives et les bruits d'explosion
   – vibration des cordes vocales pour les sons vocaliques (voyelles en particulier).
      ⇨ fréquence fondamentale 120Hz pour un locuteur, 200Hz pour une locutrice

2. Calcul d'un spectrogramme : DFT d'une fenêtre de 4 a 32 ms qui se déplace de la moitié de sa durée. Avec une fréquence d'échantillonnage de 16kHz, cette fenêtre a donc entre 64 et 512 points.
      ⇨ DFT entre 64 et 512 points
      ⇨ pour lisser le spectre on utilise en fait une DFT avec plus de points (au moins 256) ce qui permet de mieux d'interpoler le spectre.
      ⇨ "zero padding" : le signal de départ complété par des zéros. Si on utilise une DFT de 512 points et que la fenêtre a 64 points on complète par (512 - 64) zéros.

3. fenêtre plus courte que la période fondamentale (spectre a large bande)
      ⇨ fort lissage fréquentiel et pas d'harmonique

4. fenêtre plus longue (spectre a bande étroite).

⇨ faible lissage fréquentiel et visualisation des harmoniques.

Exemples de spectrogramme :

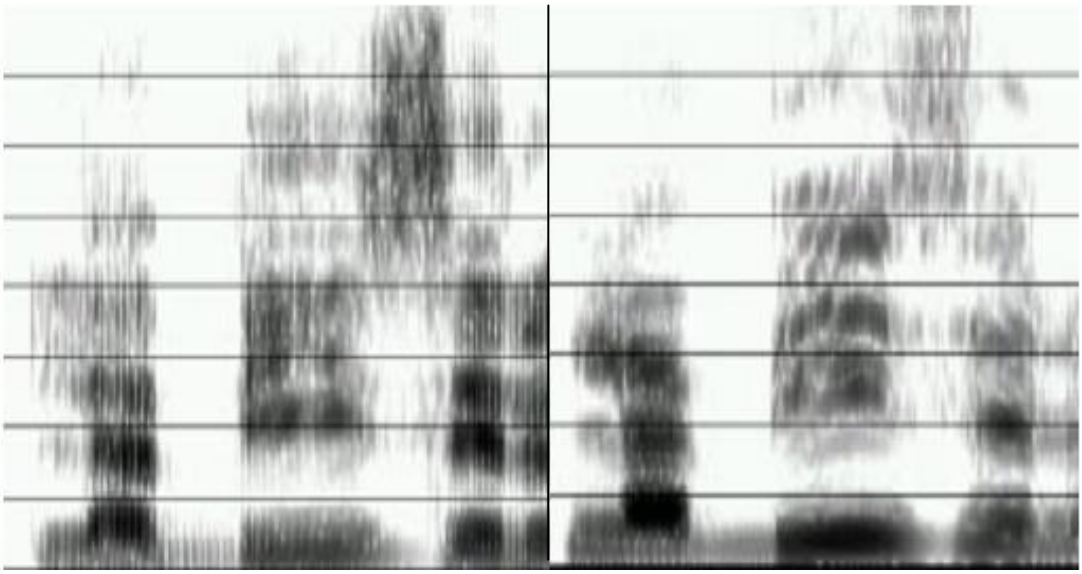


Figure IV.7 Spectrogrammes a bande large (locuteur a gauche, locutrice a droite)

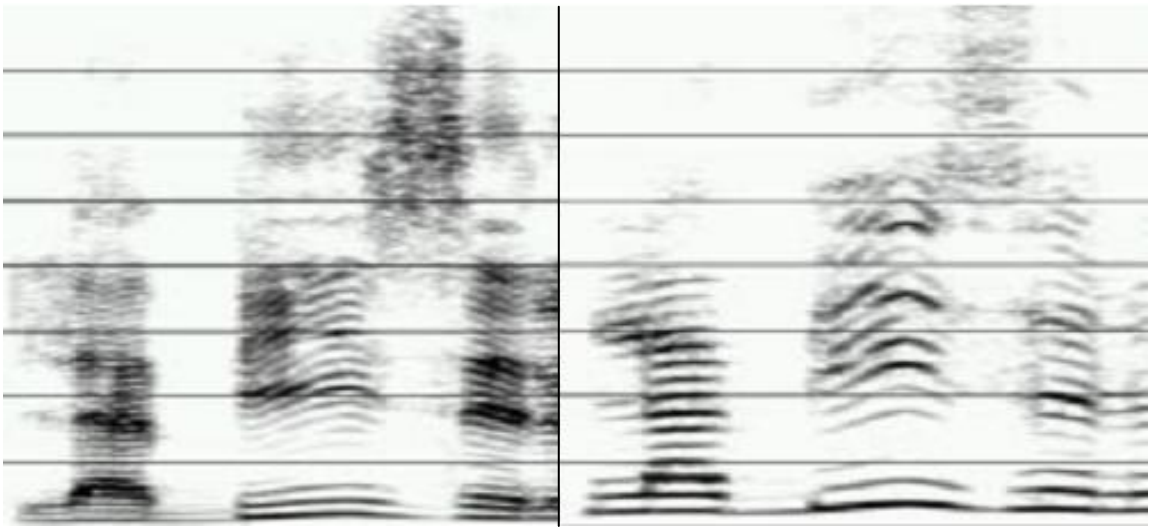


Figure IV.8 Spectrogrammes a bande étroite (locuteur a gauche, locutrice a droite)

IV.3.2 La transformée rapide de Fourrier (FFT)

En 1965, Cooley et Tukey proposèrent une méthode qui permet de réduire considérablement le temps de calcul de la TFD d'une suite dont le nombre d'échantillons N est décomposable en facteurs (typiquement, une puissance de 2).

a) La FFT radix 2 avec entrelacement dans le temps

Cette méthode, qui exige une séquence dont la longueur est une puissance de 2 ($N = 2^M$), a rendu envisageable le calcul de TFD de plusieurs milliers de points. Le nom de radix 2 provient du fait que l'on ramène le calcul d'une TFD de N points à un certain nombre de calculs de TFD de 2 points. L'appellation entrelacement dans le temps est liée à la décomposition de la suite {x(n)} en suites plus courtes.

Soit {x(k)} la TFD d'une suite {x(n)} de longueur $N = 2^M$:

$$X(k) = \sum_{n=0}^{N-1} x(n)\, W^{-nk} \qquad (k = 0 \dots N-1)$$

Soient les deux suites a(n) et b(n) de longueur N/2 et leurs TFD A(k) et B(k) :

a(n) = x (2n) a(n) ⇔A(k)

b(n) = x (2n+1) b(n) ⇔B(k)

On montre facilement que les X(k) peuvent être calculés partir des A(k) et B(k) :

$$X(k) = \sum_{n=0}^{\frac{N}{2}-1} x(2n)\, W_n^{-2nk} + \sum_{n=0}^{\frac{N}{2}-1} x(2n+1)\, W_n^{-(2n+1)k}$$

$$= A(k) + W^{-k}B(k)$$

$$X\left(\frac{N}{2}+k\right) = \sum_{n=0}^{\frac{N}{2}-1} x(2n)\, W_n^{-2n\left(k+\frac{n}{2}\right)} + \sum_{n=0}^{\frac{N}{2}-1} x(2n+1)\, W_n^{-(2n+1)\left(k+\frac{N}{2}\right)}$$

$$= A(k) - W^{-k}B(k)$$

IV.4 Conclusion

Une fois appréhendé tous les outils nécessaire au traitement du signal (parole), ils seront appliqué dans la phase préliminaire de la reconnaissance de la parole (mots isolés), a fin d'analyser le signal et en extraire les informations acoustiques après avoir choisis un modèle (stochastique) sur lequel sera basée la reconnaissance.

CHAPITRE V

La reconnaissance vocale

L'absence dans le signal vocal d'indicateurs sur les frontières de *phonèmes* et de mots constitue une difficulté majeure de la reconnaissance de la parole. De ce fait, la reconnaissance de mots prononcés artificiellement de façon isolée (c'est à dire que tous les mots prononcés sont séparés par des silences de durées supérieures à quelques dixièmes de seconde) représente une simplification notable du problème.

Deux systèmes ont cours actuellement :

- Le système mono locuteur (utilisable par un seul locuteur) est caractérisé par la technique d'apprentissage, où une seule et même personne doit dicter un ensemble de mots, ce qui permet d'optimiser le taux de reconnaissance et d'étendre le vocabulaire utilisable. Inconvénient, seule la personne ayant fourni son empreinte vocale (lors de la phase d'apprentissage) peut travailler.

- Le système multi locuteur (utilisable par plusieurs locuteurs) qui utilise une base de données contenant des empreintes moyennes autorisant la reconnaissance de plusieurs

voix. Inconvénient, le système n'est pas doté de capacités d'apprentissage et le nombre de mots est plus limité. [2]

V.1 Techniques de reconnaissance vocale

Deux approches, l'une plus globale, l'autre plus analytique permettent d'appréhender la reconnaissance des mots.

Dans l'approche globale, l'unité de base sera le plus souvent le mot considéré comme une entité globale, c'est à dire non décomposée. L'idée de cette méthode est de donner au système une image acoustique de chacun des mots qu'il devra identifier par la suite. Cette opération est faite lors de la phase d'apprentissage, où chacun des mots est prononcé une ou plusieurs fois. Cette méthode a pour avantage d'éviter les effets de coarticulation, c'est à dire l'influence réciproque des sons à l'intérieur des mots. Elle est cependant limitée aux petits vocabulaires prononcés par un nombre restreint de locuteurs.

L'approche analytique, qui tire parti de la structure linguistique des mots, tente de détecter et d'identifier les composantes élémentaires (phonèmes, syllabes, …). Celles-ci sont les unités de base à reconnaître. Cette approche a un caractère plus général que la précédente : pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base.

Pour la reconnaissance de mots isolés à grand vocabulaire, la méthode globale ne convient plus car la machine nécessiterait une mémoire et une puissance considérable pour respectivement stocker les images acoustiques de tous les mots du vocabulaire et comparer un mot inconnu à l'ensemble des mots du dictionnaire. Il est de plus impensable de faire dicter à l'utilisateur l'ensemble des mots que l'ordinateur a en mémoire. C'est donc la méthode analytique qui est utilisée : les mots ne sont pas mémorisés dans leur intégralité, mais traités en tant que suite de phonèmes.

V.2 Principe général de la méthode globale pour un système mono locuteur

Le principe de base est le même que ce soit pour l'approche analytique ou l'approche global, ce qui les distingues c'est que : pour la première il s'agit de reconnaitre le phonème, pour l'autre le mot. La structure d'un système de reconnaissance de mots isolés est représentée sur la Figure V.1. Dans l'utilisation d'un tel système, on peut distinguer deux phases:

- La phase d'apprentissage : un locuteur prononce l'ensemble du vocabulaire, souvent plusieurs fois, de façon à créer le dictionnaire de références acoustiques. Pour l'approche analytique, la machine demande à l'utilisateur d'énoncer des phrases souvent dépourvues de toute signification, mais qui présentent l'intérêt de comporter des successions de phonèmes bien particuliers. Pour un système multi locuteur, cette phase n'existe pas, c'est la principale différence.

- La phase de reconnaissance : un locuteur (le même que précédemment car nous sommes dans le cas d'un système mono locuteur) prononce un mot du vocabulaire. Ensuite la reconnaissance du mot est un problème typique de reconnaissance de formes. Tout système de reconnaissance de formes comporte toujours les trois parties suivantes:

    - Un capteur permettant d'appréhender le phénomène physique considéré (dans notre cas un microphone),
    - Un étage de paramétrisation des formes (par exemple un analyseur spectral),
    - Un étage de décision chargé de classer une forme inconnue dans l'une des catégories possibles.

On retrouve ces trois étages dans un système de reconnaissance vocale, comme le montre la figure ci-dessous.

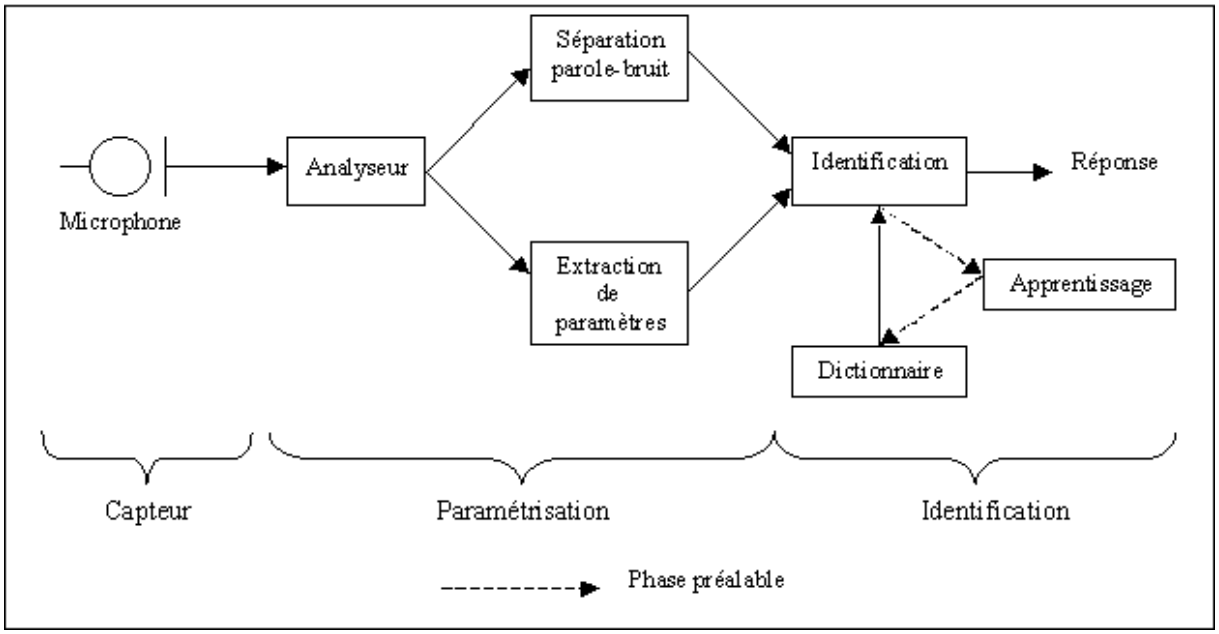


Figure V.1 Système de reconnaissance de mots isolés [2]

V.3 Paramétrisation du signal

La parole apparait comme une variation de la pression de l'air dans l'appareil phonatoire humain. Les différents traits acoustiques du signal de parole sont notamment : sa fréquence fondamentale, son énergie, son spectre. . . Chacun de ces éléments étant lui-même intimement lié a une grandeur perceptible : pitch, intensité, timbre. .etc.

L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes. Un tel système prend un signal en entrée et retourne un vecteur de paramètres (appelé indifféremment vecteur acoustique ou encore vecteur d'observations). Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs). [16]

Il existe un certain nombre d'approches pour la paramétrisation. Nous présentons ici celles utilisées le plus couramment:

- paramétrisation basée sur un modèle de production de la parole ;
- paramétrisation basée sur une analyse dans le domaine cepstral.

V.3.1 Paramétrisation basée sur un modèle de production de la parole

LPC (Linear Predictive Coefficients)

Cette approche est basée sur les connaissances expertes en production de la parole. Le conduit vocal est modélisé comme un filtre autorégressif (AR). Ceci permet d'approximer l'échantillon de l'instant n (s(n)) par une combinaison linéaire des p échantillons précédents (P étant l'ordre du modèle).

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i * s(n-i)$$

L'erreur de prédiction du modèle peut être estimée par :

$$e(n) = S(n) - \tilde{s}(n)$$

On peut donc estimer l'erreur quadratique moyenne par :

$$E_n = \sum_m e(m)^2 = \sum_m [s(m) - \sum_{i=1}^{p} a_i * s(m-i)]^2$$

Minimiser cette erreur quadratique revient à annuler les dérivées partielles de E_n par rapport aux a_i. Pour cela, plusieurs approches sont possibles (méthode de covariance, méthode d'auto-corrélation). [16]

V.3.2 Paramétrisation basée sur une analyse dans le domaine cepstral

Le signal de parole (S_n) est le résultat de la convolution entre un signal excitateur g_n (la glotte) et le conduit vocal b_n :

$$S_n = g_n * b_n$$

Le passage, par homomorphisme, dans un domaine ou l'opérateur de convolution est transformé en opérateur d'addition permet de dé-corréler les contributions de la source et du conduit du signal de parole. En pratique, l'utilisation de la transformée de Fourier donne les coefficients cepstraux :

$$\tilde{s}_n = \tilde{g}_n + \tilde{b}_n$$

Ou $\tilde{g}_n$ et $\tilde{b}_n$ sont les transposées dans le domaine fréquentiel de g_n et b_n
Plusieurs méthodes permettent d'obtenir ces coefficients :

- Grâce a une récursion depuis les coefficients LPC, ce qui donne les coefficients LPCC
- par l'utilisation d'une FFT et d'une FFT inverse ; cette technique permet de calculer les coefficients MFCC, LFCC et PLP. [16]

LPCC (Linear Predictive Cepstral Coefficients)

Cette méthode permet de calculer les coefficients LPCC directement depuis les coefficients LPC.

$$LPCC_i = -LPC_i + \sum_{k=1}^{i-1} (1 - k/i) LPC_k LPCC_{i-1}$$

Cette approche a pour but de modéliser davantage l'enveloppe du signal.

Le calcul des coefficients cepstraux (MFCC, LFCC et PLP) est souvent précédé d'une phase de préaccentuation du signal, suivie d'un fenêtrage :

La préaccentuation est définie de la manière suivante :

$$x[i] = x[i] - \alpha * x[i-1]$$

α est généralement compris entre 0,90 et 1 (la valeur classique de α est 0,97)

Le fenêtrage appliqué généralement est celui de Hamming :

$$Ham[i] = 0.54- O.46*\cos(2*\pi*i/N)$$

Ou N correspond à la longueur de la fenêtre.
Ces deux étapes sont préalables au calcul des coefficients MFCC, LFCC et PLP. [16]

MFCC (Mel Frequency Cepstral Coefficient)

Afin de rapprocher l'analyse en banc de filtres de la perception humaine, les filtres ne sont généralement pas répartis de manière linéaire mais en fonction d'une échelle Mel. La correspondance entre une fréquence en Hz et en Mel se calcule de la manière suivante :

$$F_{mel} = 2595*\text{Log}\ (1+ (F_{Hz}/700))$$

Intuitivement, cela revient à utiliser une échelle linéaire en basse fréquence, puis logarithmique en haute fréquence. Généralement, seuls les 12 premiers coefficients cepstraux sont conservés et une vingtaine de filtres sont utilisés pour l'analyse en banc de filtres.

La chaine complète de calculs des coefficients MFCC est définie par la figure suivante :

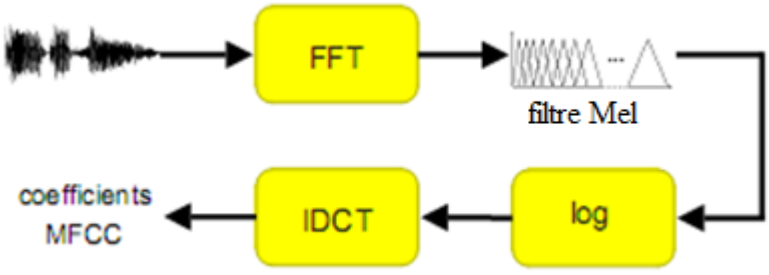


Figure V.2 Chaine de traitement pour obtenir les coefficients MFCC. [16]

LFCC (Linear Frequency Cepstral Coefficient)

Il s'agit d'une variante des MFCC. La différence vient de l'utilisation d'un banc de filtres linéaire, contrairement a l'échelle Mel des MFCC.

PLP (Perceptual Linear Predictive)

Cette extraction de paramètres est basée sur des connaissances expertes de l'appareil auditif humain.

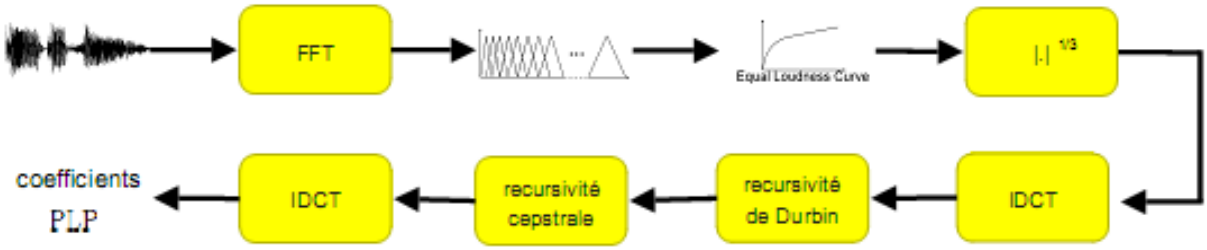


Figure V.3 Chaine de traitement pour obtenir les coefficients PLP [16]

La paramétrisation du signal en coefficients PLP est finalement assez proche d'une LPC.

V. 4 Principe de décodage

Les premières recherches en reconnaissance automatique de la parole ont débuté à partir des années 1950. Dans les années 1960, une méthode appelée Dynamic Time Warping (DTW) est apparue. Elle repose sur les travaux de [Bellman 1957] et reste aujourd'hui une approche importante en reconnaissance de mots isolés.

Une seconde méthode a émergé dans les années 1975, avec les travaux de [Jelinek 1976]. Elle s'appuie sur l'utilisation des modèles de Markov cachés (Hidden Markov Models - HMM). Elle a permis de nombreuses avancées dans les domaines de la reconnaissance de la parole continue et de la reconnaissance multi-locuteurs, domaines dans lesquels la DTW était peu probante.

V.4.1 Déformation dynamique temporelle (DTW)

Principe général :

L'idée directrice de la DTW consiste à estimer une mesure de similarité entre la représentation d'un mot référence et la représentation d'un mot inconnu afin d'évaluer l'écart entre ces deux mots.

Pour cela, nous disposons d'un ensemble de références R_x qui forment le dictionnaire (C) des n mots à reconnaitre : $C = \{R_x\}_{1 \leq x \leq n}$. Muni d'une distance D, il devient alors possible de calculer un cout de déformation entre le mot inconnu (T) et une référence x (R_x). Le but est donc de réaliser un alignement temporel, le meilleur qui soit, entre une référence et un mot à tester. Le mot prononcé est trouvé par la résolution de :

$$t = \text{Argmin } D(T,R_x)$$

Soient R_x la référence d'un mot du dictionnaire et T le mot à reconnaitre de longueurs respectives I et J. R est composée d'éléments r(1), r(2), r(3), ...,r(I) (respectivement pour T : t(1), t(2), t(3), ..., t(J)) qui représentent les vecteurs acoustiques du signal à un instant donné.
On appelle d(ri,tj) la distance entre les vecteurs acoustiques r(i) et t(j). La figure suivante illustre ce principe.

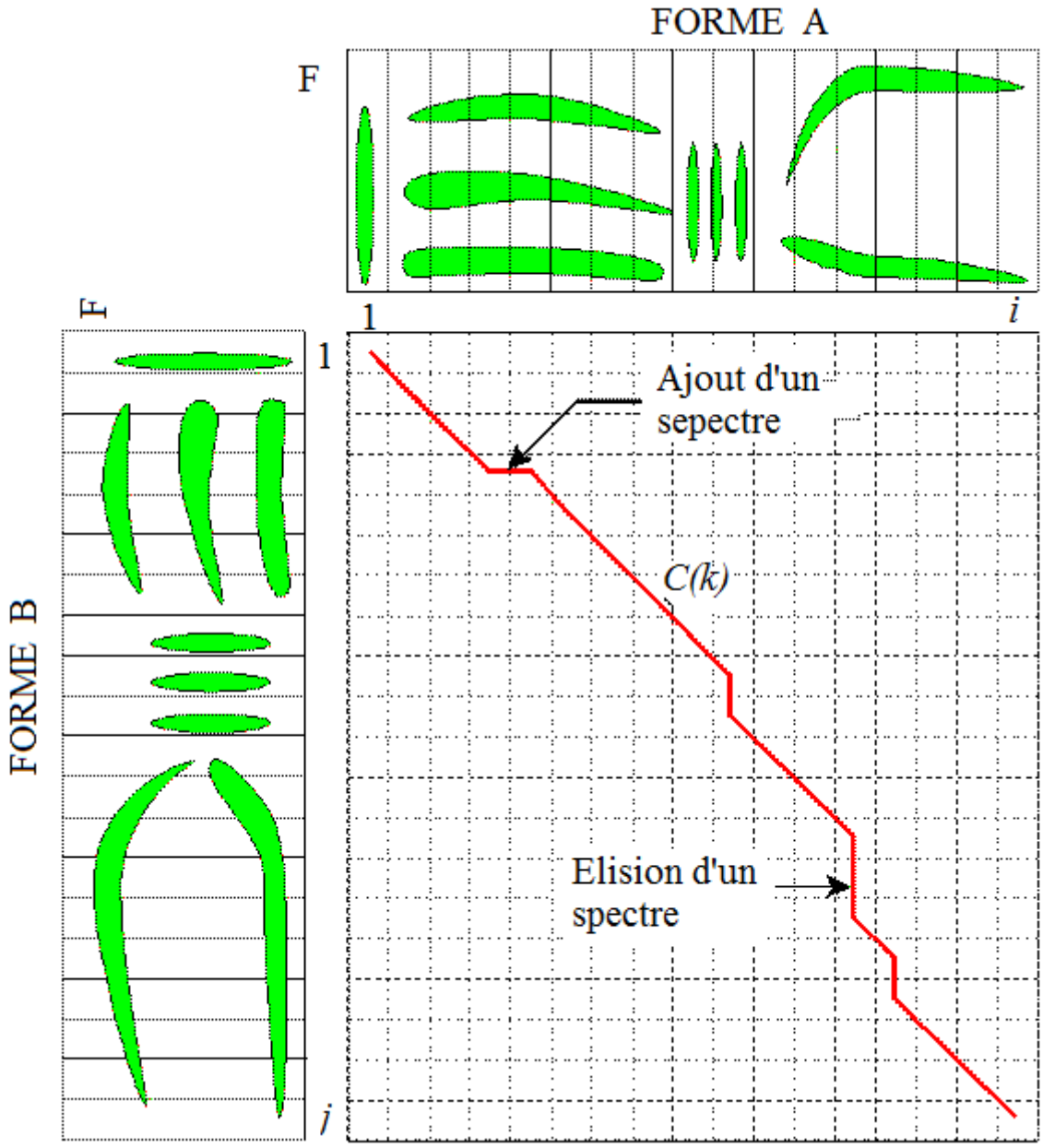


Figure V.4 Principe de la DTW [5]

Distances usuelles :

Plusieurs distances peuvent être utilisées en fonction des méthodes de paramétrisation retenue :

- La distance utilisant les normes L_n (n allant de 1 ∞) est plutôt utilisée dans les systèmes a base d'analyse cepstrale. La norme L_2 est la plus utilisée. Elle est plus connue sous le nom de distance euclidienne :

$$d_2(r_i, t_j) = (\sum_{k=0}^{p} |r_k - t_k|^2)^{1/2}$$

- La mesure d'Itakura est plutôt utilisée dans le cadre d'une paramétrisation par prédiction linéaire

$$d_{it}(r_i, t_j) = \log [\frac{r_i^t R_b r_i}{t_i^t R_b t_j}]$$

Avec R_b représentant la matrice des coefficients d'auto corrélation évalués sur le segment t_j.

Contraintes locales :

Afin de tenir compte des réalités physiques du mécanisme de production de la parole, les déplacements entre les vecteurs de paramètres sont limités (contraintes locales). Les contraintes locales les plus courantes sont représentées dans la figure suivante.

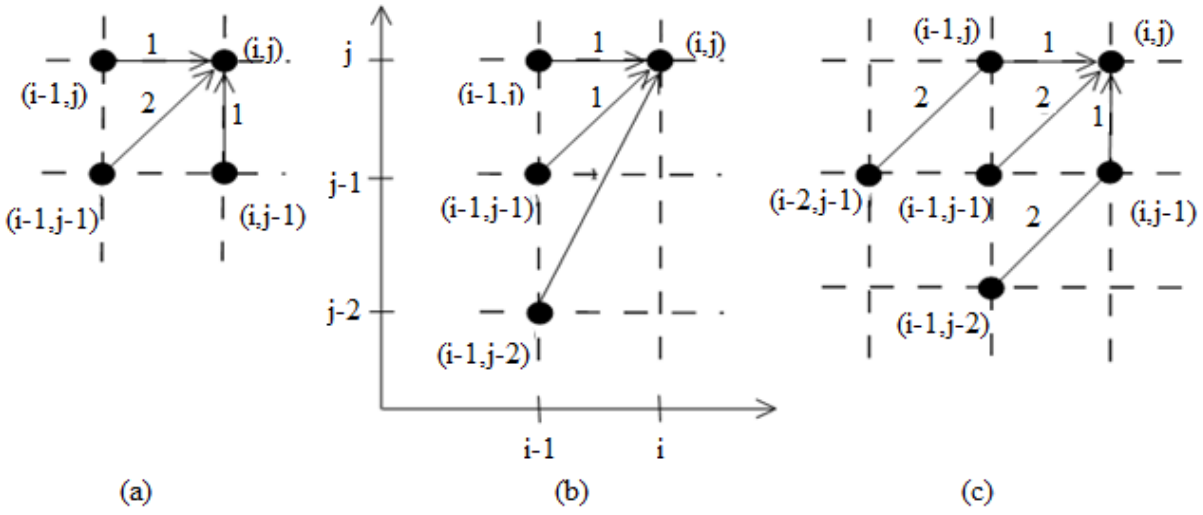


Figure V.5 Contraintes locales utilisées dans la DTW [5]

Le principe est donc de trouver le chemin d'alignement ayant un cout minimum. La distance cumulée g (i,j) est définie (en fonction de la contrainte locale choisie) par :

$$g(i,j) = \min \begin{cases} g(i-1, j) + d(i, j) \\ g(i-1, j-1) + 2*d(i, j) \\ g(I, j-1) + d(i, j) \end{cases}$$

En normalisant par les longueurs de R et T, on obtient donc :

$$D(R, T) = g(I, J)/ (I+J)$$

V.4.2 Modèles de Markov cachés (HMM)

La reconnaissance automatique de la parole, vue comme un problème de la théorie de communication, a pour but de reconstruire un message M inconnu à partir d'une séquence d'observations O. Ceci revient à retrouver, parmi tous les messages possibles, celui qui selon toute vraisemblance, correspond à la suite d'observations acoustiques O. Cette dernière correspond à une suite de vecteurs permettant de caractériser le signal de parole. Le message M est une suite de mots prononcés. Il s'agit donc de trouver M qui correspond au message le plus probable connaissant la suite des observations acoustiques O. [9]

$$\widehat{M} = arg_M max P(M\backslash O)$$

La probabilité P(M\O) est très difficile à déterminer, d'où la nécessité de la décomposer. En utilisant la règle de Bayes, il est possible de reformuler la probabilité P(M\O) comme suit :

$$P(M\backslash O) = \frac{P(M)P(O\backslash M)}{P(O)}$$

Puisque P(O) ne dépend pas de M, l'équation précédente sera équivalente à :

$$\widehat{M} = arg_M maxP(M)P(O\backslash M)$$

Ainsi, l'étape de reconnaissance consiste à déterminer la suite de mots $\widehat{M}$ qui maximise le produit des deux termes P(M) et P(O\M). Le premier terme représente la probabilité a priori d'observer la suite de mots M indépendamment du signal. Cette probabilité est déterminée par le modèle de langage. Le deuxième terme indique la probabilité d'observer la séquence de vecteurs acoustiques O sachant une séquence de mots spécifiques M. Cette probabilité est estimée par le modèle acoustique.

La Figure V.6 montre les différentes étapes nécessaires à la reconnaissance d'un message m prononcé en entrée. Tout d'abord, le signal de parole est subdivisé en vecteurs acoustiques. En utilisant ces vecteurs, le modèle acoustique se charge, à partir des HMM de phonèmes appris sur un corpus d'apprentissage, de construire la suite des phonèmes hypothèses du signal prononcé. Un seul modèle HMM représentant l'hypothèse, sera construit par la concaténation d'un ensemble de HMM de phonèmes. La suite de mots obtenue sera aussi évaluée par le modèle de langage qui permet d'estimer la probabilité P(M). En principe, ce processus est répété pour toutes les hypothèses possibles. Le système donne les N meilleures hypothèses comme résultat de la reconnaissance. [9]

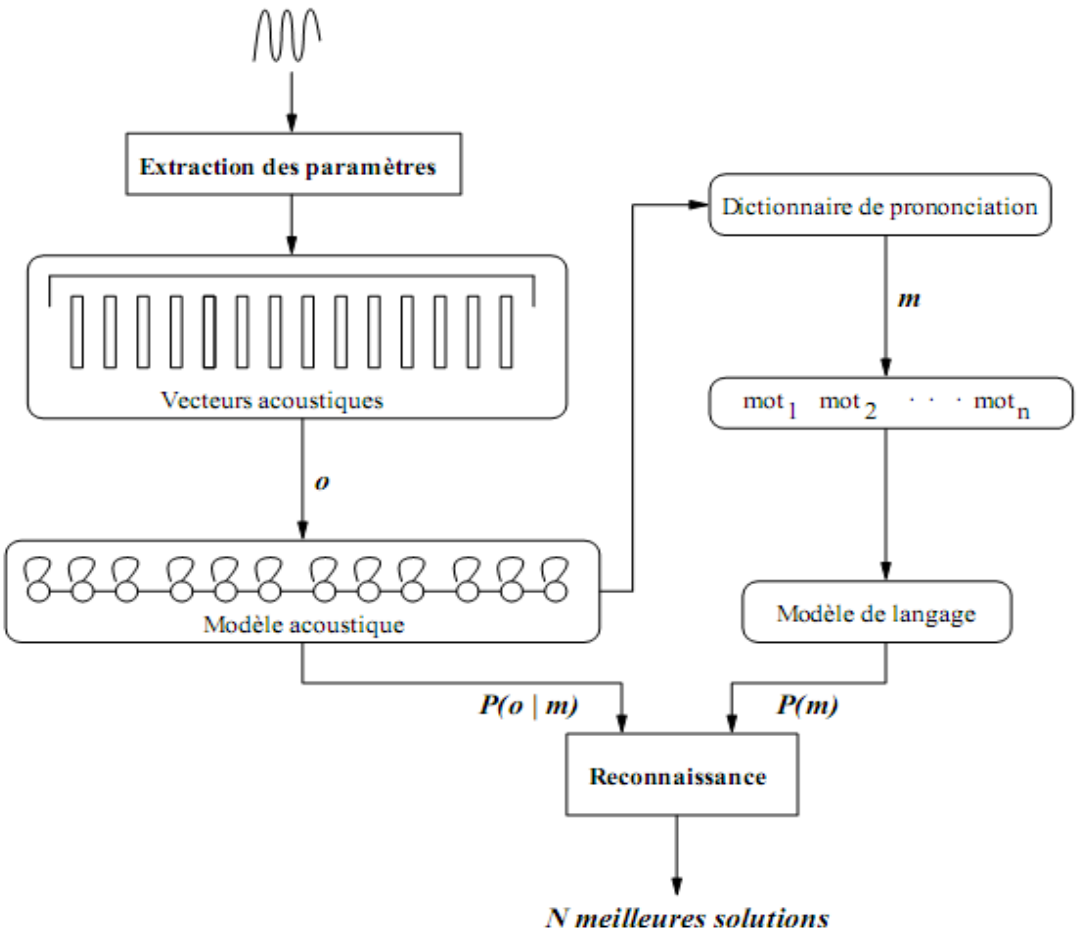


Figure V.6 Le principe de fonctionnement de l'approche statistique pour la reconnaissance automatique de la parole. [9]

V.4.2.1 Modélisation acoustique à base de HMM

Un HMM peut être vu comme un ensemble discret d'états et de transitions entre ces états. Formellement, il peut être défini par l'ensemble des paramètres λ.

$\lambda = (N, A, B, \pi)$

Où :

– - N est le nombre de nœuds ou d'états du modèle.

– -A = $\{a_{ij}\}$ = {P ($q_j \backslash q_i$)} est une matrice de taille N*N. Elle contient les probabilités de transition sur l'ensemble des états du modèle. La probabilité de transition est la probabilité de choisir la transition $a_{i\,j}$ pour accéder à l'état q_j en partant de l'état q_i.

-B = $\{b_j(o_t)\}$ = $\{P(o_t \backslash q_j)\}$, où j appartient à l'intervalle [1, N], est l'ensemble des probabilités d'émission de l'observation o_t sachant qu'on est dans l'état q_j. La forme que prend cette distribution détermine le type du HMM.

-π est la distribution initiale des états, $\pi_j = P(q_0 = j)$, (j :1, N). q_0 représente l'état initial du modèle HMM. Il ne peut émettre de vecteurs acoustiques. [9]

V.4.2.2 Les problèmes fondamentaux des HMM

Évaluation :

Étant donnée une séquence d'observations O = $o_1, o_2, \ldots o_T$ et le modèle λ = (N, A, B, π), la question qui se pose est : comment calculer efficacement $P(O \backslash \lambda)$, la probabilité d'observer la séquence O sachant le modèle λ ?

Pour calculer cette probabilité il faut tout d'abord, définir la probabilité d'observer la séquence O pour une séquence d'états Q = $q_1, q_2, \ldots q_T$:

$$P(O \backslash Q, \lambda) = \prod_{t=1}^{T} b_{q_t}(o_t)$$

Or, la probabilité de la séquence Q peut s'écrire sous la forme suivante :

$$P(Q \backslash \lambda) = \pi_{q1} \prod_{t=2}^{T} b_{q_{t-1} q_t}$$

La probabilité conjointe du chemin Q et des observations O est donnée par :

$$P(O, Q \backslash \lambda) = P(Q \backslash \lambda) P(O \backslash Q, \lambda)$$

La probabilité de la séquence d'observations O sachant le modèle λ est obtenue par :

$$P(O\backslash\lambda) = \sum_{Q} P(O, Q\backslash\lambda)$$

Il vient alors :

$$P(O\backslash\lambda) = \sum_{q_1,q_2,\ldots,q_T} \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} b_{q_{t-1}q_t}\, b_{q_t}(o_t)$$

Le calcul direct de cette probabilité nécessite beaucoup de calculs, ce qui nous mène à utiliser l'algorithme « avant-arrière » (Forward-Backward) :

- L'algorithme avant-arrière :

Soit $\alpha_t(i)$ la probabilité d'observer o_1, o_2,..o_t et d'aboutir a l'état $q_t=S_i$, et $\beta_t(i)$ la probabilité d'observer o_{t+1}, .o_T sachant que l'on part de l'état $q_t=S_i$.

$$P(O\backslash\lambda) = \sum \alpha_t(i) * \beta_t(i).$$

-Calcul des fonctions α :

Il est réalisé par l'algorithme suivant :

Pour i allant de 1 a N faire

$\alpha_1(i) = \pi_i b_i(o_1)$

Pour t allant de 1 a T-1 faire

Pour j allant de 1 a N faire

$$\alpha_{t+1}(j) = \left[\sum a_{ij} * \alpha_t(i)\right] * b_j(o_{t+1})$$

$P(O\backslash\lambda) = \sum \alpha_T(i)$

-*Calcul des fonction β:*

Il est par l'algorithme suivant :

Pour i allant de 1a N faire

$\beta_T(i) = 1$;

Pour t allant de T-1 a 0 faire

Pour i allant de 1 a N faire

$\beta_t(i) = \sum a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$.

$p(O\backslash\lambda) = \sum \pi_i \beta_0(j)$. [5]

Décodage:

Étant donnés une séquence d'observations $O = o_1, o_2 \ldots o_T$ et le modèle $\lambda = (N, A, B, \pi)$, comment choisir la séquence d'états $Q = q_1, q_2 \ldots q_T$ qui a le plus de chance d'émettre la séquence d'observations O?

En effet, le problème de décodage revient à chercher une séquence d'états « optimale », Q*. Cela peut être fait de différentes façons. La principale difficulté réside dans la définition de la séquence d'états optimale. Il faut donc commencer par choisir un critère parmi plusieurs critères d'optimalité possibles. Le critère le plus utilisé est celui qui cherche la meilleure séquence d'états globale (le meilleur chemin)

$$Q^* = arg \max_Q P(O, Q\backslash\lambda)$$

Ceci nous ramène :

$$P(Q, O\backslash\lambda) = \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} b_{q_{t-1}q_t}\, b_{q_t}(o_t)$$

De même que pour le problème d'évaluation, le calcul direct nécessite plusieurs opérations. On utilise l'algorithme de Viterbi :

On représente le meilleur chemin aboutissant à q_T ; elle peut être calculée par induction

$\omega_{t+1} = (max\ [\omega_t(i)]*a_{ij})*b_j(o_{t+1})$

Soit la fonction $\psi_{t+1}(i)$ qui représente l'état correspondant au chemin optimal :

$\psi_{t+1}(i) = arg\ max\ [\omega_t(i)*a_{ij}]$

Pour i allant de 1 a N faire

$\omega_1(i)=\pi_i*b_i(o_1)$

$\psi_1(i)= 0$

Pour t allant de 1 a T faire :

Pour j allant de 1 a N faire

$\omega_t(j) = max_i\ [\omega_{t-1}(i)*a_{ij}]*b_j(o_t)$

$\psi_{t(j)} = arg\ max_i\ [\omega_{t-1}(i)*a_{ij}]$

$P^* = max_i\ [\omega_T(i)]$

$q_T^* = arg\ max_i\ [\omega_T(i)]$

Pour t allant T-1 a 1 faire

$q_t^* = \psi_{t+1}(q_{t+1}^*)$ [5]

Apprentissage :

Comment déterminer les paramètres du modèle λ = (N, A, B, π) afin de maximiser P(O\λ) ?

Autrement dit, le problème d'apprentissage consiste à trouver une méthode qui permet d'ajuster les paramètres du modèle λ = (N, A, B, π) et de maximiser la probabilité d'une séquence d'observations donnée, sachant le modèle λ. Formellement, il s'agit de déterminer A*, B* et π* tels que :

$$(A^*, B^*, \pi^*) = \arg \max_{(A,B,\pi)} P(O\backslash A, B, \pi)$$

Ce problème n'a pas de solution analytique connue et il n'existe pas de technique optimale pour estimer les paramètres du modèle. On peut cependant choisir $\lambda = (N, A, B, \pi)$ de telle façon que $P(O\backslash\lambda)$ soit localement maximale.

La méthode généralement utilisée dans ce genre de cas, est une procédure itérative qui vise à optimiser le critère de maximum de vraisemblance sur un corpus d'apprentissage donné. Il s'agit de l'algorithme EM (Expectation-Maximization)

Algorithme EM :

Soit la probabilité de passage d'un état S_i a un état Sj au temps t

$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j / O,\lambda)$

Alors :

$\xi_t(i,j) = (\alpha_t(i)*a_{ij}*bj(o_{t+1})*\beta_{t+1(j)})/(P(O/\lambda))$

$= (\alpha_t(i)*a_{ij}*b_j(o_{t+1})*\beta_{t+1}(j))/ \sum_i\sum_j \alpha_T(i)*a_{ij}*b_j(o_{t+1})*\beta_{t+1}(j)$

Soit la probabilité de se trouver dans l'état S_i au temps t

$\gamma_T(i) = P(q_T = S_i / O,\lambda)$

on a en particulier $\gamma_T(i) = \sum_j \xi_t(i,j)$

Estimation des paramètres :

$\pi^*_i = \gamma_1(i) = \alpha_1(i)*\beta_1(i)/\sum_i \alpha_t(i)*\beta_t(i)$

$a^*_{ij} = [\sum_{t=1}^{T-1} \xi_t(i,j)] / [[\sum_{t=1}^{T-1} \gamma_T(i)]$

$b^*_j(k) = = [\sum_{t=1\ t:\{t/ot=vk)}^{T-1} \xi_t(i,j)] / [[\sum_{t=1}^{T-1} \gamma_T(i)]$ [5]

V.4.2.3 Application des HMM dans la reconnaissance de la parole

En reconnaissance de la parole, des modèles de Markov gauche-droite d'ordre un sont le plus souvent utilisés du fait de l'aspect séquentiel du signal de parole (modèle de Bakis).

La figure suivante illustre un exemple d'un HMM à 3 états typique utilisé pour la modélisation d'un phonème. Les états d'entrée et de sortie sont ajoutés pour faciliter la concaténation des modèles entre eux. L'état de sortie d'un modèle de phonème peut être fusionné avec l'état d'entrée d'un autre modèle de Markov caché pour former un modèle composite.

Ceci permet aux modèles de phonèmes d'être concaténés ensemble pour former les mots et ainsi les phrases. On remarque que les seules transitions permises sont de type gauche-droite,

dans le but de mieux modéliser la contrainte temporelle de la parole. Un HMM est considéré comme un générateur de vecteurs acoustiques, c'est une machine à états finis qui change d'état à chaque unité de temps. Pour chaque unité de temps t, une fois arrivé à l'état q_j, un vecteur acoustique o_t est généré avec une densité de probabilité $b_j(o_t)$. De plus, la transition de l'état q_i à l'état q_j est probabiliste, sa probabilité est généralement notée a_{ij}. En pratique, c'est seulement la séquence d'observations $O = o_1, o_2, \ldots o_T$ qui est connue. La séquence d'états est non observable directement, d'où le nom de modèle de Markov « caché ». [9]

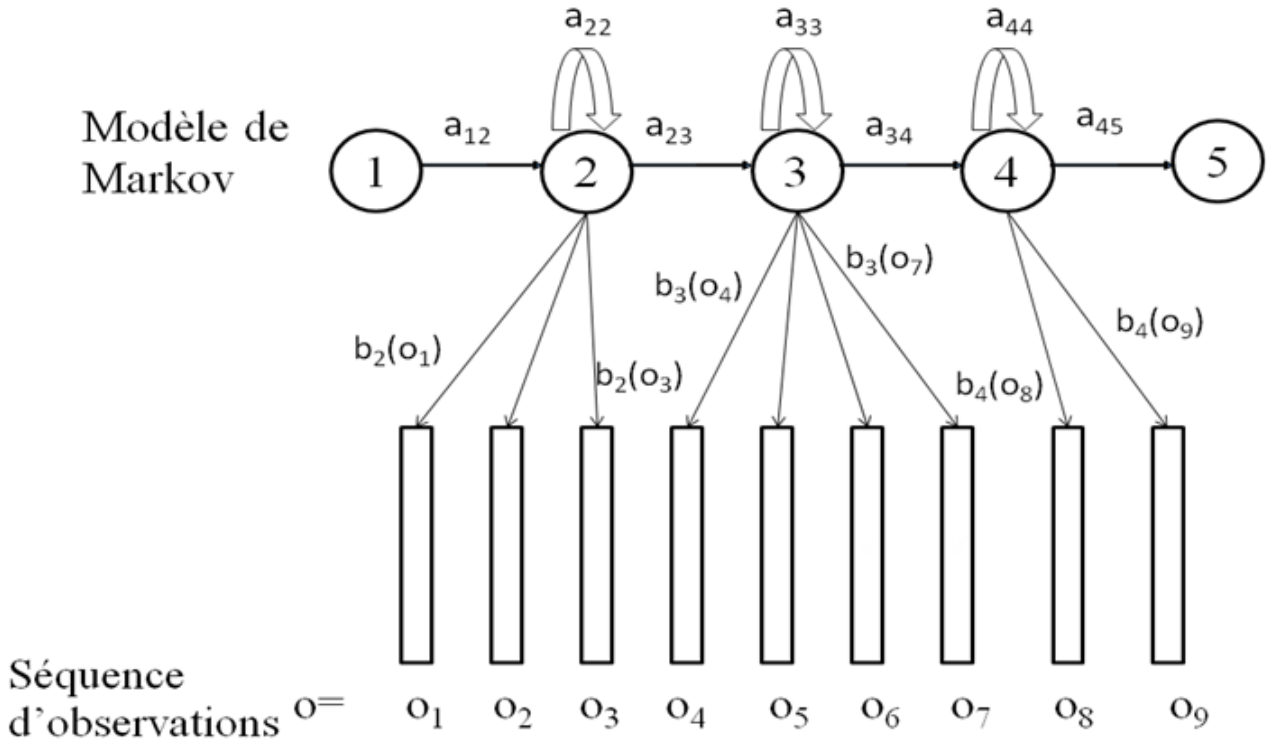


Figure V.7 Un modèle HMM à 5 états gauche-droite [14]

L'approche phonétique pour la reconnaissance de la parole consiste à ne modéliser que les phonèmes. Elle permet ainsi, d'éviter la collecte d'énormes corpus d'apprentissage qui sont nécessaires pour la modélisation des HMM de mots. Ceci rend réalisable la reconnaissance automatique de la parole sur de grands vocabulaires.

La combinaison de plusieurs HMM de phonèmes en un seul est illustrée par la FigureV.8, où un HMM correspondant au mot Paris est obtenu par la concaténation des HMM correspondant aux phonèmes **p, a, r, i.** La Figure V.8, montre que la construction peut être

obtenue par la coïncidence du dernier état du HMM d'un phonème avec le premier état du HMM du phonème suivant :

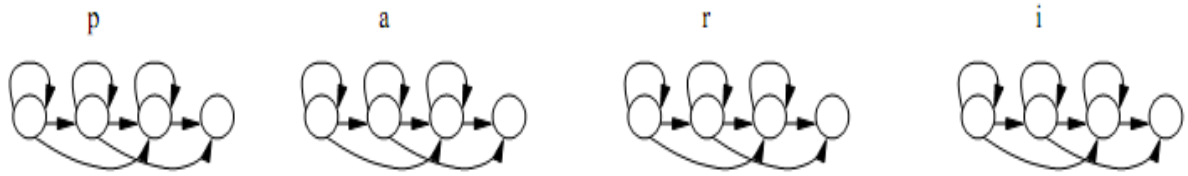


Figure V.8 HMM des phonèmes p,a,r,i [9]

Un HMM plus élaboré est illustré par la Figure V.9, où pour le même mot, on ajoute une transition qui permet l'omission d'un phonème. Finalement, de la même façon qu'on peut construire des HMM de mots à partir de HMM de phonèmes, on peut construire des HMM de phrases à partir de HMM de mots comme le montre la Figure V.10.

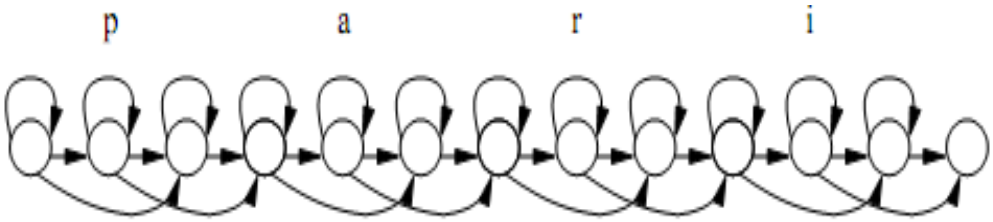


Figure V.9 HMM du mot ‘pari’ [9]

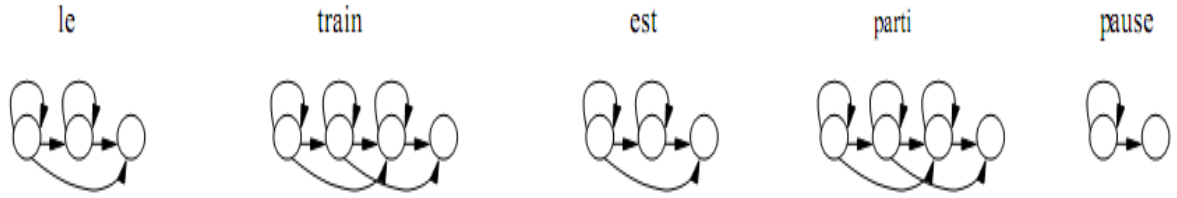


Figure V.10 : HMM d’une phrase (‘le train est parti’) [9]

V.4.3 Le modèle du langage

La modélisation acoustique permet à elle seule de réaliser la transcription phonétique d'une phrase. Cependant, en l'absence de contraintes d'ordre linguistique, qu'elles soient lexicales, syntaxiques ou sémantiques, la suite de phonèmes obtenue peut s'avérer très lointaine de la chaîne attendue. En effet, l'auditeur humain exploite les niveaux supérieurs pour lever l'ambiguïté phonétique et même pour compenser une dégradation limitée de l'information acoustique. De plus, une reconnaissance acoustique même parfaite ne suffit pas pour obtenir une transcription correcte de la phrase. En effet, une suite de 9 phonèmes peut être transcrite en français en 32000 suites de mots différentes orthographiquement correctes, mais dont seulement quelques unes correspondent à des phrases syntaxiquement correctes. Il est donc indispensable d'introduire dans le système de reconnaissance des connaissances d'ordre linguistique afin d'améliorer ses performances. [9]

Un modèle de langage probabiliste est construit à partir d'un très grand corpus d'apprentissage composé de données exprimées dans la langue étudiée. Il a pour but d'estimer la probabilité P(W) où W est une suite de mots W = w1, w2, . . . wN. Notons que le modèle de langage doit pouvoir évaluer cette probabilité même si la suite des mots W n'a pas été rencontrée dans le corpus d'apprentissage.

V.4.3.1 Les modèles n-grammes

Grâce à leur simplicité et à leur efficacité, les modèles n-grammes constituent les modèles de langage les plus employés dans le domaine de la reconnaissance de la parole. Ils sont basés sur l'hypothèse que l'apparition d'un mot dépend seulement de son historique proche. Détaillons donc cette méthode de modélisation du langage.
En appliquant la règle des probabilités conditionnelles, la vraisemblance de la suite de mots W = w1, w2, . . . wN est définie comme suit :

$$P(w_1, w_2, \dots, w_N) = \prod_{i=1}^{N} p(w_i \backslash w_1, \dots, w_{i-1})$$

où $p(w_i \backslash w_1, \ldots, w_{i-1})$ est la probabilité du mot w_i sachant tout les mots qui le précédent et qui constituent son historique.

Les valeurs de n les plus utilisées sont n = 2 et n = 3 et on parle dans ces deux cas respectivement, de modèles *bigrammes* et de modèles *trigrammes*. Pour n = 1, nous parlons du modèle *unigramme* qui ne tient pas compte de l'historique des mots. [5]

V.4.3.2 Les modèles n-classes

En raison du manque de données d'apprentissage, il est nécessaire de trouver une méthode qui permet de maximiser la quantité d'information utile et ce en réduisant l'espace des paramètres du modèle. Une des méthodes retenues consiste à regrouper les mots en classes. En effet, avec un nombre de classes inférieur au nombre de mots du vocabulaire, il y aura beaucoup moins d'évènements à modéliser. Par conséquent, le modèle aura une plus grande capacité de généralisation. Le regroupement des mots en classes peut être fait selon plusieurs critères. Nous pouvons trouver dans la littérature, trois types de classes : les classes syntaxiques qui regroupent les mots selon leur catégorie grammaticale, les classes morphologiques qui regroupent les mots ayant la même racine morphologique (lemme) dans une seule classe et enfin les classes obtenues par d'autres méthodes de classification automatique.

Les modèles n-classes consistent à attribuer à chaque mot w_i une classe $C(w_i)$ et à estimer les probabilités des mots en fonction de deux facteurs : la probabilité d'appartenance du mot à sa classe $P(w_i \backslash C(w_i))$ et la probabilité de l'apparition de cette classe à la suite de son historique de classes. Dans le cas où un mot ne peut appartenir qu'à une seule classe, la probabilité du mot w_i sachant son historique $w_{i-n+1} \ldots w_{i-1}$ est définie comme suit :

$$p(w_i \backslash w_{i-n+1}, \ldots, w_{i-1}) = p(w_i \backslash C(w_i)) * p(C(w_i) \backslash C(w_{i-n+1}), \ldots, C(w_{i-1}))$$

Le terme $p(C(w_i) \backslash C(w_{i-n+1}) \ldots C(w_i\text{-}_1))$ correspond à la probabilité de la succession des classes $C(w_{i-n+1}) \ldots C(w_{i-1})C(w_i)$ est estimé de la même manière que la probabilité $p(w_i \backslash w_{i-n+1} \ldots w_{i-1})$

La probabilité d'appartenance du mot w_i à la classe $C(w_i)$, quant à elle, est calculée selon la formule suivante :

$$p(w_i \backslash C(w_i)) = \frac{N(w_i)}{N(C(w_i))}$$

où N est le nombre d'occurrences de l'argument dans le corpus et C est la classe à la quelle appartient le mot en argument. [5]

V.5 Reconnaissance de mots isolés (évaluation expérimentale)

L'objectif de notre expérience est de pouvoir reconnaitre 9 chiffres (de 1 jusqu'à 9) en utilisant les modèles de Markov cachés.

V.5.1 Présentation de la base de données

Notre base de données est constituée de deux grandes parties :

La partie utilisée pour l'apprentissage contient 171 séquences audio, réparties sur 9 chiffres prononcés par 19 locuteurs différents en termes d'âge et de sexe.

La partie utilisée pour les tests contient 108 séquences, réparties de la même manière que la précédente.

Ces séquences audio ont été enregistrées via un microphone intégré dans téléphone mobile Sony-Eriksson K770i en format *.amr* ensuite converties en *.wav* sur un micro-ordinateur pour pouvoir les charger sur *Matlab*.

V.5.2 L'apprentissage

Cette étape consiste à construire les modèles HMM (un modèle par mot), en utilisant les fonctions *apprenti.m* (ex : apprenti1 correspond au premier mot) qui utilisent l'algorithme de *maximum de vraisemblance* EM, qui nous retournent les paramètres suivants :

- Prior : qui représente la matrice π, (les probabilités initiales du modèle) ;
- Transmat : qui représente la matrice A, (les probabilités de transition du modèle).
- Mixmat et Sigma : qui représentent la matrice B, (les probabilités d'émission des observations).

A la fin des calculs on sauvegarde ces données dans le *Workspace* en format *.MAT* en les organisant de la manière suivante :

-*dataTR1* : les paramètres du premier modèle.
-*dataTR2* : les paramètres du deuxième modèle.

-*dataTR9* : les paramètres du neuvième modèle.

V.5.3 Reconnaissance et évaluation

On commence par charger le fichier à reconnaitre (le chiffre à tester en format *.wav*) comme entrée de la fonction *Test.m* tout en chargeant tous les fichiers réalisés durant l'apprentissage (*dataTR*). Cette fonction utilise l'algorithme de maximum de vraisemblance qui va chercher le meilleur modèle qui correspond au mot qui a été chargé qui a obtenu le meilleur score. Le résultat est visualisé comme graphe qui contient les valeurs du logarithme de la probabilité de chaque modèle (*Loglik*).

Exemple 1 :

On veut reconnaitre le mot *wordtest11* qui correspond au mot un prononcé par le premier locuteur.

1- On charge le fichier (en utilisant la commande *wavread*) dans la fonction Test.m

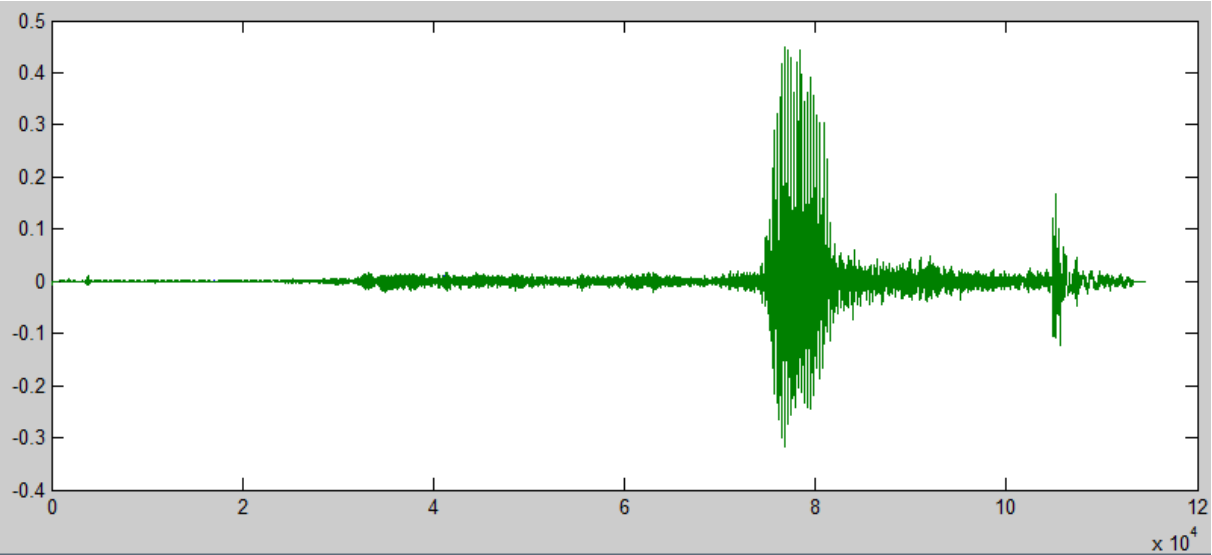


Figure V.11 Audiogramme du mot « wordtest11 »

2- On visualise les résultats de cette fonction

Le résultat est le suivant :

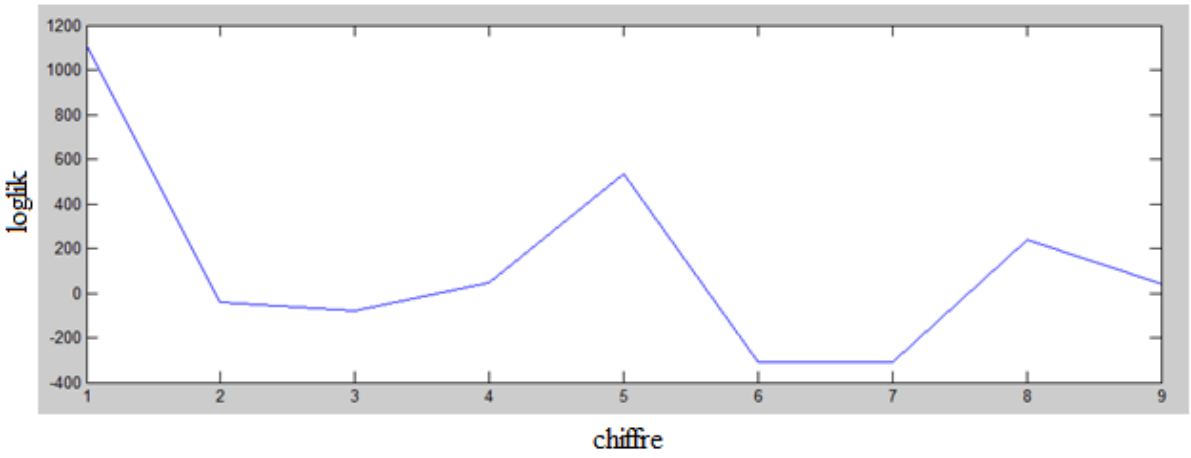


Figure V.12 Réponse de la reconnaissance 1

On constate que l'amplitude maximale correspond au chiffre « 1 *: un* » et qui correspond au chiffre prononcé dans la séquence.

Exemple 2 :

On essaie un deuxième mot *wordtest24,* qui correspond au mot «2 : deux » prononcé par le quatrième locuteur. On refait les mêmes étapes et on visualise le résultat suivant :

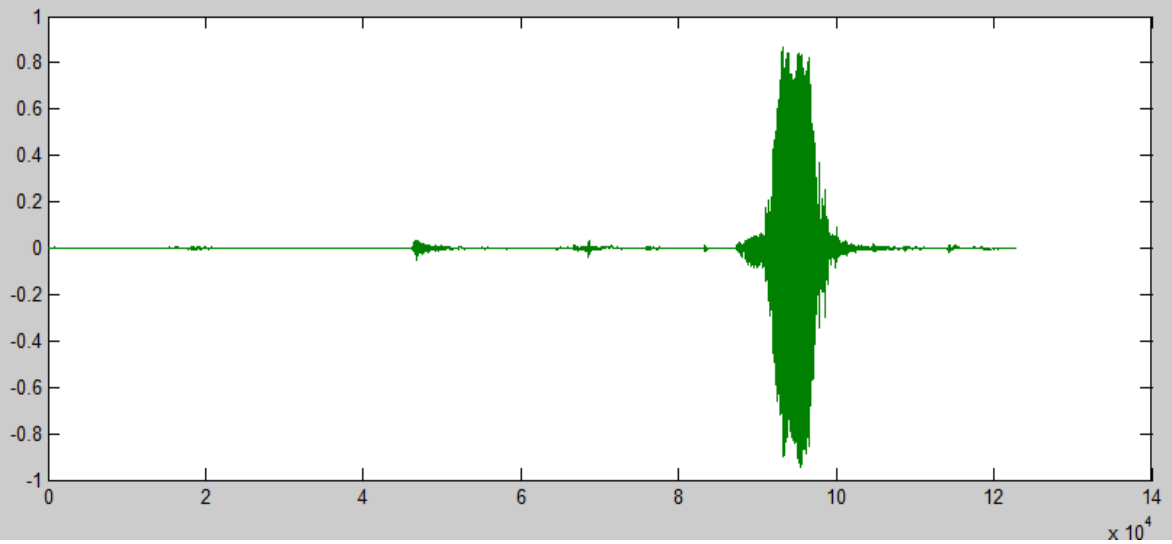


Figure V.13 Audiogramme du mot « wordtest24 »

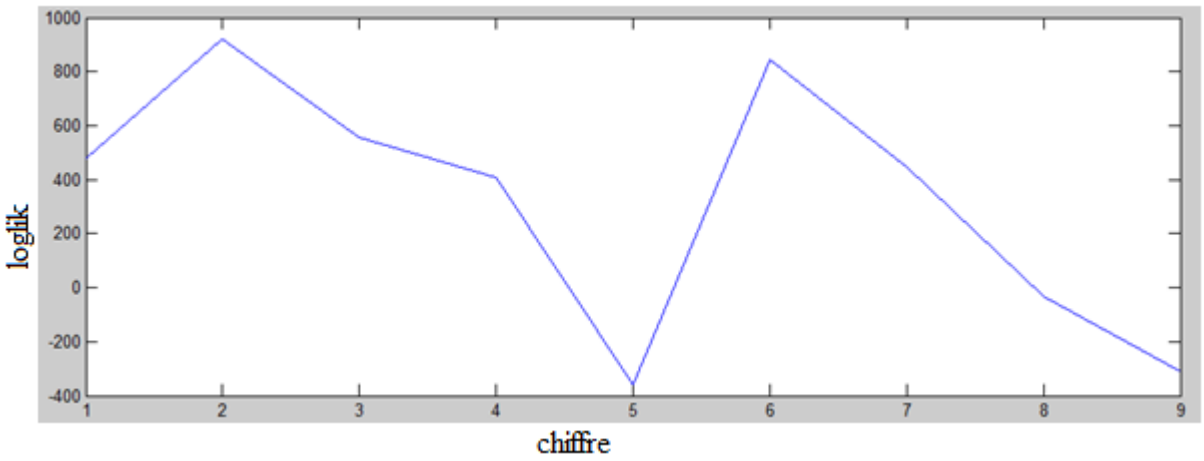


Figure V.14 Réponse de la reconnaissance 2

L'amplitude maximum correspond au chiffre « *2 : deux* », qui correspond au chiffre qui a été prononcé dans la séquence chargée.

V.5.4 Taux de reconnaissance :

Nous nous proposons d'illustré ci-dessous des tableaux, les différents taux de reconnaissance qu'on a obtenu, en faisant varier le nombre de séquences durant la phase d'apprentissage, ceci est présenté sous forme de 4 tableaux :

- pour 5 séquences d'apprentissage :
  Nous avons obtenu les résultats suivant

Mot a reconnaitre	Nombre de mots reconnus	Taux de reconnaissance/mot
Un	2/12	16.16%
Deux	3/12	25%
Trois	6/12	50%
Quatre	8/12	66.66%
Cinq	4/12	33.33%
Six	5/12	41.66%
Sept	7/12	58.33%
Huit	6/12	50%
Neuf	6/12	50%
Taux de reconnaissance globale : 43.51%		

Tableau V.1 Taux de reconnaissance avec 5 séquences d'apprentissage

- pour 10 séquences d'apprentissage :
  Nous avons trouvé les résultats suivants

Mot a reconnaitre	Nombre de mots reconnus	Taux de reconnaissance/mot
Un	3/12	25%
Deux	5/12	41.66%
Trois	9/12	75%
Quatre	8/12	66.66%
Cinq	6/12	50%
Six	9/12	75%
Sept	8/12	66.66%
Huit	5/12	41.66%
Neuf	9/12	75%
Taux de reconnaissance globale : 65.74%		

Tableau V.2 Taux de reconnaissance avec 10 séquences d'apprentissage

- pour 15 séquences d'apprentissage :

  Nous avons trouvé les résultats suivants

Mot a reconnaitre	Nombre de mots reconnus	Taux de reconnaissance/mot
Un	10/12	83.33%
Deux	12/12	100%
Trois	11/12	91.11%
Quatre	7/12	58.33%
Cinq	9/12	75%
Six	9/12	75%
Sept	12/12	100%
Huit	10/12	83.33%
Neuf	11/12	91.11%
Taux de reconnaissance globale : 86.11%		

Tableau V.3 Taux de reconnaissance avec 15 séquences d'apprentissage

- pour 19 séquences d'apprentissage :

  Nous avons trouvé les résultats suivants

Mot a reconnaitre	Nombre de mots reconnus	Taux de reconnaissance/mot
Un	12/12	100%
Deux	12/12	100%
Trois	12/12	100%
Quatre	10/12	83.33%
Cinq	12/12	100%
Six	9/12	75%
Sept	11/12	91.66%
Huit	10/12	83.33%
Neuf	8/12	66.66%
Taux de reconnaissance globale : 88.88%		

Tableau V.4 Taux de reconnaissance avec 19 séquences d'apprentissage

Interprétations : On constate que :

- Pour le même nombre de séquences d'apprentissage, on a des taux de reconnaissance différents par mot :
  - La représentation phonétique est différente d'un mot à un autre (pour le *un* qu'un seul phonème alors que le *quatre* à quatre phonèmes) ce qui engendre des variations de prononciations d'un locuteur à un autre.
  - Des prononciations différentes engendrent des difficultés de reconnaissance ce qui donne des taux différents.
  - Mauvaise conditions d'enregistrement (miro a faible directivité) dans des salles bruitées.

- Des taux de reconnaissance différents en fonction du nombre de séquences d'apprentissage :
  - L'algorithme de ré-estimation dispose a chaque fois de plus de données.
  - Construction de modèles de plus en plus précis ce qui rend le système de plus en plus performant.

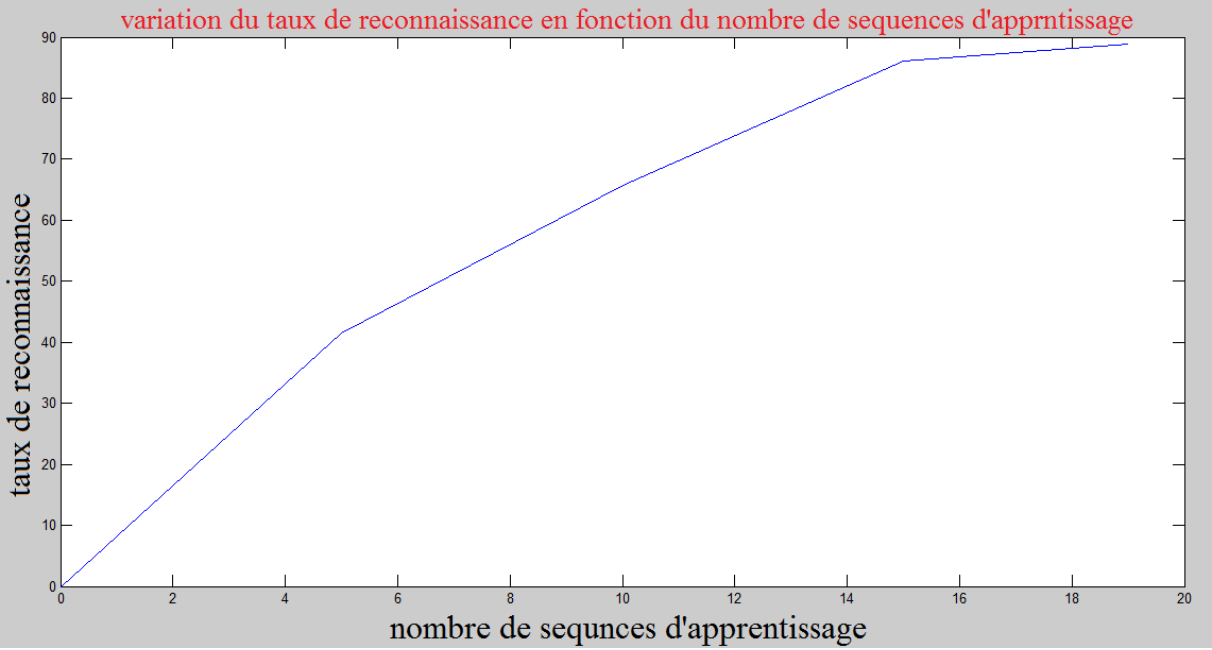


Figure V.15 Variation du taux de reconnaissance en fonction du nombre de séquences d'apprentissage

V.6 Conclusion

Après avoir étudié les méthodes stochastiques, en particulier les Modèles de Markov cachés (HMM), nous l'avons utilisé pour faire un système de reconnaissance de mot isolé avec une paramétrisation MFCC, et nous avons obtenu d'assez bon résultats, tout fois nous avons utilisé un vocabulaire restreint, avec un nombre de locuteurs limité. Pour la réalisation

d'un système de reconnaissance avec un très grand vocabulaire, il faudra faire un apprentissage beaucoup plus performant (une centaine d'essais) et de bonnes conditions d'enregistrement et préparation de la base de données à l'étude (faire l'enregistrement dans des salles phonétiquement isolées et élimination des bruits de fond) pour pouvoir avoir un taux de reconnaissance global acceptable.

CHAPITRE VI

L'analyse lexico-syntaxique

Dans cette partie nous allons nous intéressons à la compréhension du langage parlé, pour ce faire, après la phase de reconnaissance nous étudierons ces résultats pour une analyse lexicale.

L'élaboration de systèmes de compréhension de parole a pour but de faciliter le dialogue homme-machine. Il est donc logique d'y intégrer un module de dialogue fondé sur des informations pragmatiques propres à l'application traitée.

VI.1 La composante lexico-syntaxique

VI.1.1 Le Lexique

VI.1.1.1 Introduction

Depuis des années, les questions lexicales ont pris une importance considérable. Le lexique le lien entre les différents niveaux d'un système complet de reconnaissance

automatique de la parole, à savoir : les niveaux acoustique, phonétique, phonologique, prosodique, morphologique, syntaxique, sémantique et pragmatique.

Il est centre d'information pour les traitements utilisés dans ces systèmes.

A cet effet, le lexique doit avoir une meilleure représentation des différents niveaux (3) :

-la première catégorie appelée: traitement bas niveau, contient les niveaux acoustiques, phonétiques, phonologiques et prosodiques.

-la deuxième catégorie appelée : traitement haut niveaux, contient quand à elle les niveaux syntaxique, sémantique et pragmatique.

On peut illustrer cela par la Figure VI.6, et qui représente le niveau lexical en compréhension de la parole. [5]

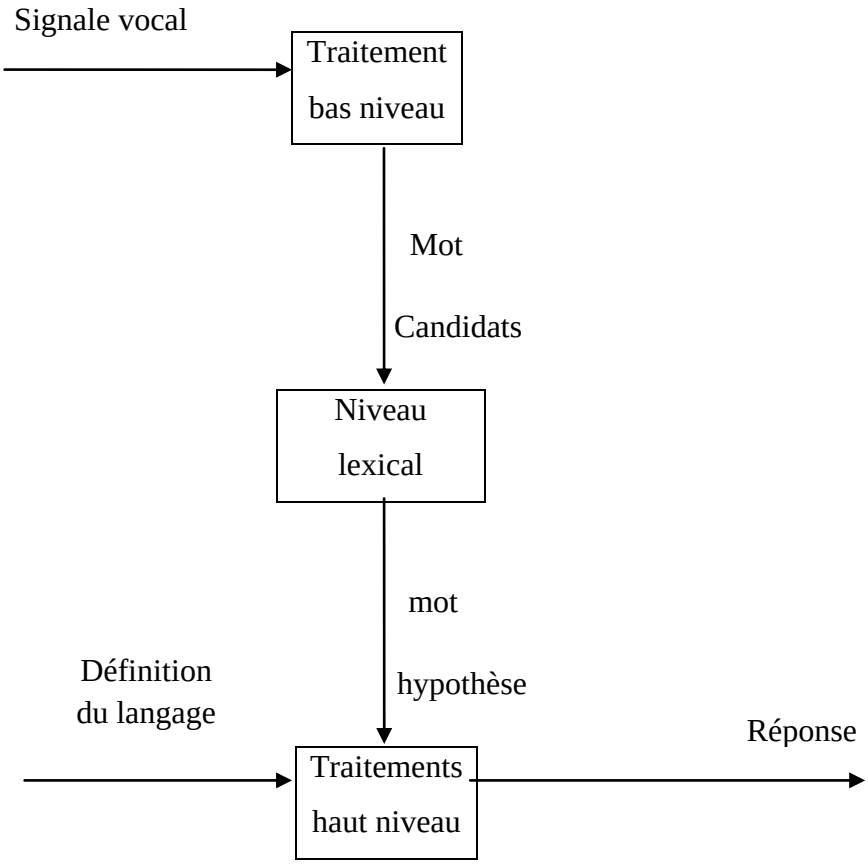


Figure VI.6 Rôle du niveau lexical en compréhension de la parole

Vu que notre travail ne traite que les composantes lexicales et syntaxiques. Nous n'allons utiliser que les résultats du traitement bas niveau, ce qui fait que les paramètres utilisés pour ce dernier ne seront représentés que par la transcription phonétique des mots.

VI.1.1.2 Organisation de la composante lexicale :

a) bases de données lexicales :

Intégrer un corpus lexical important, en vue de l'utiliser dans un traitement informatique, pose de véritables problèmes de taille mémoire, de complexité de structure de gestion et d'accès aux données.

De plus, on doit envisager la possibilité de partager le lexique entre plusieurs utilisateurs ou plusieurs applications.

Pour se faire, nous avons opté pour un lexique dynamique. La saisie des mots utilisés pour une application se fait indépendamment des autres applications, avec le choix de garder ou pas l'ancien lexique. [5]

b) Nature des entrées lexicales :

Il s'agit de déterminer le type d'élément qui va constituer le lexique. Pour se faire, une idée des différentes possibilités qui existent, on peut examiner les formes d'organisation utilisées dans les dictionnaires destinés à la distribution grand public, et dans les lexiques à vocation généralement plus linguistique. Il existe deux conceptions différentes pour le lexique, celle des lexicographes et celle des lexicologues. [5]

b-1) Organisation lexicographique :

Cette organisation propose des lexiques comportant des faits de tout ordre, destinés à être consultés par des usagers, locuteurs d'une langue mais non spécialistes de celle-ci.

Les mots qui composent la langue française sont divisés en deux catégories, les mots grammaticaux et les mots lexicaux.

Les premiers sont limités par leur nombre et ne posent pas de problèmes d'organisation, ils se présentent sous une seule forme, c'est le cas des pronoms en général. Par contre les mots lexicaux sont les noms, les adverbes, les adjectifs et les verbes. A l'exception des adverbes invariables, tous les mots lexicaux possèdent les formes fléchies qui correspondent pour les noms et les adjectifs à la flexion en genre et en nombre, et pour les verbes à la conjugaison.

Forme fléchie :

Un tel type d'organisation implique que toutes les formes propres à un mot de la langue étudiée, soient représentées dans le lexique, ce qui équivaut à disposer d'un lexique en extension. Ceci présente un certain intérêt dans la mesure où l'on est dispensé de la mise en œuvre des processus flexionnels permettant d'obtenir les formes fléchies. Mais surtout, cela permet de traiter de manière simple les exceptions et les formes présentant des irrégularités.

Un tel schéma d'organisation n'est convenable que pour une langue comme l'anglais qui représente un nombre réduit de formes fléchies, par opposition aux langues latines.

Une simple comparaison du nombre de formes propres à un verbe anglais et à un verbe française suffit à illustrer ce propos.

Le nombre de formes fléchies en français est environs dix fois plus grand qu'en anglais.

Forme canonique :

Ce mode d'organisation, est utilisé par la plus part des dictionnaires, cela consiste à ne représenter les formes verbales, nominales ou adjectivales que sous un seul aspect. Il s'agit de focaliser les formes fléchies d'un même mot modulo la flexion genre/nombre ou la conjugaison et d'en faire une classe d'équivalence qui sera représentée dans le lexique par une forme.

b-2) Organisation lexicologique :

Le modèle linguistique du fonctionnement du langage, puis proposé un modèle plus orienté lexique qu'autre chose.

Cette organisation présente ainsi des formes différentes suivant l'unité lexicale choisie. Certains linguistes sont partisans d'une morphologie du mot, d'autre d'une morphologie du morphème.

a-Morphologie du mot :

Dans ce cas, on considère que l'unité Morphologique doit être le mot.

b-morphologie du morphème :

Le morphème est une unité morphologique inférieure au mot, il possède, une partie plus ou moins étendue des propriétés du mot, à l'exception des traits syntaxique.

b-3) Unité lexicale :

L'unité lexicale est la composante de base dans un système de reconnaissance automatique de la parole continue. Il faut donc choisir la plus proche de l'unité de traitement dans de tels systèmes. Par ailleurs, on se ramène toujours à la reconnaissance des mots.

En gros l'unité lexicale doit alors véhiculer toutes les informations, nécessaires au traitement. Vu la complexité de ce type de traitement, on peut se poser la question de savoir s'il faut considérer le mot sous son aspect fléchi ou pas.

Parmi les représentations citées (canonique et fléchie), nous avons choisi une organisation basée sur la forme canonique complétée par un système de génération de formes fléchies. [5]

VI.1.2 La syntaxe

VI.1.2.1 Introduction

Le niveau syntaxique joue un rôle important dans un système de reconnaissance automatique de la parole. La syntaxe définit les séquences des mots autorisées. Elle permet de limiter le nombre de solutions qui peuvent interpréter le treillis phonétique. Son objectif est donc d'extraire parmi l'ensemble des combinaisons possibles de mot, le sous ensemble de solutions syntaxiques admissibles, les modèles syntaxiques issus des travaux en linguistique dans le cadre des langues écrites, nous avons opté pour le mode des grammaires génératives de Chomsky.

VI.1.2.2 Grammaire générative

Ce type de grammaire se présente comme une suite de règle de réécriture. Le mécanisme de dérivation obtenu en appliquant, les unes après les autres, ces différentes règles permettent d'engendrer l'ensemble de phrases syntaxiquement correctes du langage. Nous illustrons deux méthodes d'analyse.

a) Analyse syntaxique descendante :

Elle consiste à appliquer un algorithme qui, partant de l'axiome, construit un chemin de proche en proche en suivant les règles de production de la grammaire sur une phrase.

Exemple : soit G une grammaire dont les règles de production sont :

<PHRASE>=<GN><GV> (1)

<GN>=Pronoms personnels (2)

<GV>=Verbe (3)

{PHRASE, GN, GV} : ensemble du vocabulaire terminal.

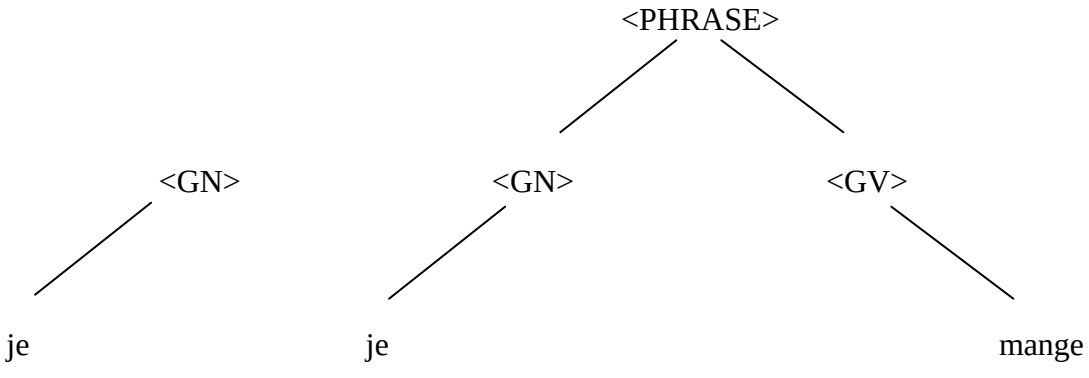
Soit à analyser la phrase suivante : « je mange » l'analyse descendante de cette phrase se fait comme suit :

<PHRASE> ➔ <GN><GV> ➔ je<GV> ➔ je mange

b) Analyse syntaxique ascendante :

Cette analyse construit le chemin vers l'axiome à part de la phrase elle-même.

Reprenons l'exemple précédent.



VI.2.2.3 Définition de la grammaire choisie

Dans un système de reconnaissance automatique de la parole, on travaille avec des phrases énoncées. Ces phrases sont soumises à certaines règles grammaticales qui les rendent compréhensibles.

Nous avons choisi de travailler sur des phrases simples telles que :

-les phrases formées par un sujet, un verbe et un complément.

-les phrases formées par un verbe, un sujet et un complément.

Pour se fait nous avons opté pour la grammaire suivante :

<PHRASE>=<GN><GV><GC> (1)

<GV><GN1><GC1> (2)

<FORME><GN1><GV><GC> (3)

<GN>=<ART><ADJQ><NOM> (4)

<ADJNQ><NOM><ADJQ> (5)

<PRPRS>(6)

<NOM><CONJ><NOM> (7)

<ART>=un /une/des/le . . . (8)

<ADJNQ> = adjectifs possessif/adjectifs démonstratifs . . . (9)

<GN1>=<PRPRS> (10)

<GC1>=<infinitif> (11)

<infinitif><GV> (12)

<PRPRS>=pronoms personnels (13)

<NOM>=les noms (14)

<ADJQ>=adjectifs qualificatifs (15)

<GV>=verbes (16)

<GC>=<ART><NOM> (17)

<ART2><NOM> (18)

<PREP><NOM> (19)

<ART2>=au/à la/aux . . . (20)

<PREP>à/avant/contre . . . (21)

<CONJ>=avec/et . . . (22)

VI.2 Simulation de l'analyse lexico-syntaxique

Une fois la reconnaissance est faite (les chiffres 1 jusqu'à 9), le module acoustique fournit les mots candidats reconnus (succession de mots) au module linguistique qui va vérifier par la suite, la juxtaposition de ces derniers et les valider par la suite comme phrase cohérente.

VI.2.1 Présentation du langage de programmation

Cette partie d'étude relève de l'expertise, ce qui nous a poussé a utiliser le langage de programmation PROLOG (turbo-Prolog dans notre cas) qui est un langage de programmation logique convenable a notre application qui consistera a définir certaines grammaires (grammaires formelles) qui vont régir la succession de nos mots reconnus, autrement dit le système acceptera *uniquement* les successions qui répondent a ces grammaires.

VI.2.2 Les grammaires choisies

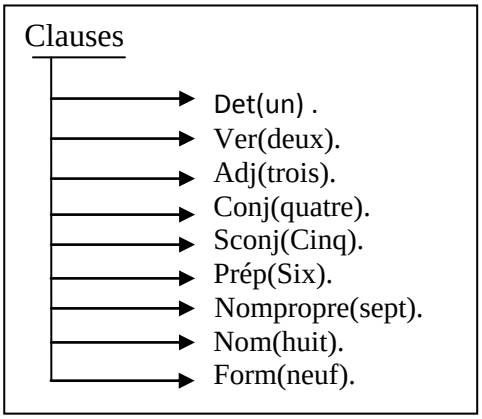
Pour se rapprocher du langage naturel, on va dire que nos chiffres reconnus correspondent à des classes syntaxiques bien précises qui sont les suivantes :

Le chiffre reconnu	*La classe syntaxique correspondante*
Un	Déterminant
Deux	Verbe
Trois	Adjectif
Quatre	Conjonction
Cinq	Conjonction de subordonnée
Six	Préposition
Sept	Nom propre
Huit	Nom
Neuf	Forme interrogative

Tableau VI.1 La classe syntaxique correspondante aux différents chiffres

On définit dans un premier temps les prédicats nécessaires aux grammaires qu'on veut définir qui sont : « *Det* », « *Ver* », « *Ad j* », « *Conj* », « *Sconj* », « *Prep* », « *Nompropre* », « *Nom* », « *Form* » qui seront déclarés comme PREDICATS en turbo-PROLOG.

L'étape suivante consiste a déclaré les clauses de la manière suivante :



Maintenant on va définir les grammaires qui vont régir les successions de mots reconnus par le module acoustique en utilisant les clauses qu'on vient de définir qui seront séparées en fonction des arguments d'entrée, qui seront vues comme des règles de production en turbo-PROLOG.

Grammaire 1 :

Det	Nom	Adj
(Un)	(Huit)	(Trois)

Exemple de langage naturel : « Le sac noir ».

Grammaire 2

Nompropre	Ver	Prep	Nom
(Sept)	(Deux)	(Six)	(Huit)

Exemple de langage naturel « JHON parle au voisin ».

Grammaire 3

Nompropre	Ver	Det	Nom
(Sept)	(Deux)	(Un)	(Huit)

Exemple de langage naturel « JHON joue du piano ».

Grammaire 4

Nompropre	Ver	Prep	Nompropre
(Sept)	(Deux)	(Six)	(Sept)

Exemple de langage naturel « JHON parle à BIL ».

Grammaire 5

Det	Nom	Ver	Det	Nom
(Un)	(Huit)	(Deux)	(Un)	(Huit)

Exemple de langage naturel « Le voisin parle le chinois ».

Grammaire 6

Det	Nom	Ver	Prep	Nompropre

(Un)	(Huit)	(Deux)	(Six)	(Sept)

Exemple de langage « Le voisin parle à JHON ».

Grammaire 7

Form	Det	Nom	Ver	Adj
(Neuf)	(Un)	(Huit)	(Deux)	(Trois)

Exemple de langage naturel « est ce que, le sac est noir ? ».

Grammaire 8

Nompropre	Conj	Nompropre	Ver	Det	Nom
(Sept)	(Quatre)	(Sept)	(Deux)	(Un)	(Huit)

Exemple de langage naturel « JHON et BIL jouent du piano ».

Grammaire 9

Nompropre	Conj	Nompropre	Ver	Prep	Nom
(Sept)	(Quatre)	(Sept)	(Deux)	(Six)	(Huit)

Exemple de langage naturel « JHON et BIL parlent au voisin ».

Grammaire 10

Det	Nom	Sconj	Nompropre	Ver	Ver	Adj
(Un)	(Huit)	(Cinq)	(Sept)	(Deux)	(Deux)	(Trois)

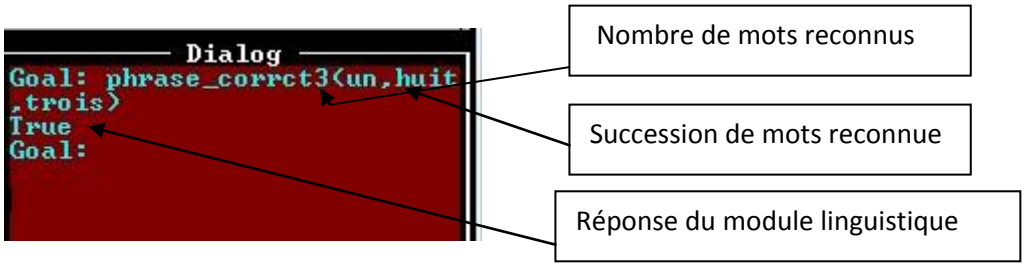
Exemple de langage naturel : « le sac que JHON a acheté est noir ».

Ces grammaires seront définies comme des clauses en turbo-PROLOG.

VI.2.3 Evaluation et simulation de l'analyse lexico-syntaxique

Le module linguistique va vérifier l'exactitude de la succession des mots reconnue de la manière suivante :

Succession de mots reconnus par le module linguistique : Un, Huit, Trois.



Succession de mots reconnus par le module linguistique : Sept, Deux, Un, Huit.

Succession de mots reconnus par le module linguistique : Sept, Quatre, Sept, Deux, Six, Huit.

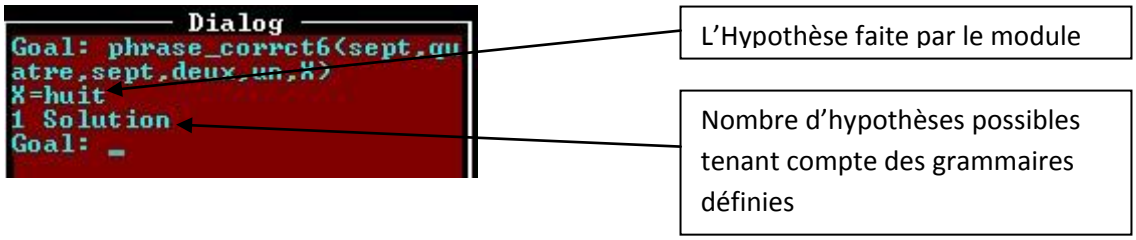
Si le module acoustique reconnait une succession de mots qui ne suit pas la grammaire définie par le module linguistique, ce dernier ne va pas la valider en répondant par un « FALSE » au lieu d'un « TRUE ».

Exemple :

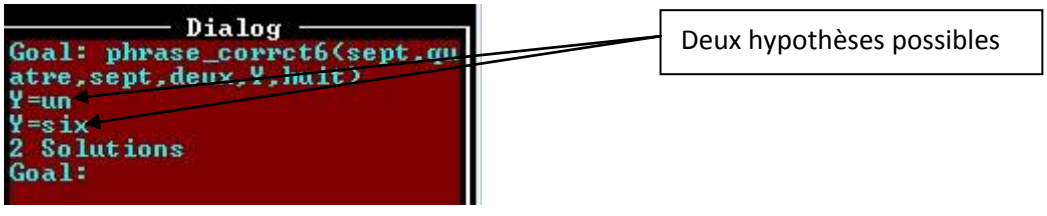
Suite de mots reconnue par le module linguistique : Sept, Quatre, Sept, Huit, Deux, Huit.

Si le module acoustique ne reconnait pas entièrement toute la séquence de chiffre, autrement dis les données disponible dans le module linguistique sont insuffisantes, ce dernier sera obligé d'émettre des hypothèses selon les possibilités qui peuvent satisfaire ses grammaires.

Exemple : a) si la séquence de mots reconnue est : Sept, Quatre, Sept, Deux, Un, *' Inconnu '* (le dernier chiffre n'est pas bien reconnu) alors le système va le poser comme inconnu X.



b) si la séquence reconnue est : Sept, Quatre, Sept, Deux, *Inconnu*, Huit.



VI.3 Conclusion

Le module linguistique (analyse lexico-syntaxique) est très important dans un système de reconnaissance de la parole, il permet de réduire le nombre de séquences de mots prononcés possibles suivant certaines grammaires (définies au préalable).

Conclusion générale et perspectives

Au cours de ce travail, nous avons traité le problème de la reconnaissance de mots isolés en mode dépendant du texte. Il s'agit d'extraire les vecteurs acoustiques à partir des signaux de parole prononcés par chaque locuteur de notre base de données, qui servent a l'entrainement (apprentissage) des modèles de chaque mot (de chaque chiffre dans notre base de données). Nous avons utilisé les paramétrés MFCC pour caractériser les signaux acoustiques en utilisant des modèles statistiques qui sont les HMM.

Nous avons montré que le taux de reconnaissance est acceptable vis-à-vis de notre application (système de reconnaissance avec vocabulaire limité), toute fois pour un système avec un riche vocabulaire demandera une phase d'apprentissage beaucoup plus robuste (une centaine de locuteurs pour construire chaque mot) pour avoir un taux de reconnaissance satisfaisant.

Nous avons ensuite montré l'importance du module linguistique qui se traduit par l'analyse lexico-syntaxique dans la compréhension automatique de la parole en définissant des grammaires inspirées d'une langue bien précise, cela pour se rapprocher du mieux de la réalité pratique en validant uniquement les suites de séquences qui vérifient ces grammaires.

Plusieurs points peuvent faire l'objet d'améliorations notables. Nous pouvons utiliser des coefficients acoustiques qui caractérisent le système auditif (modèle de l'oreille gamma-chirp) plutôt de ceux qui caractérisent le système phonatoire. Nous pouvons aussi utiliser, pour la modélisation et la construction de chaque modèle, une approche hybride statistique/connexionniste. Elle réduit l'espace mémoire nécessaire ainsi la sensibilité aux bruits.

Bibliographie

[1] A.V.Oppenheim, R.W.Shaffer, Digital signal processing. Prentice Hall, New Jersey, 1975.

[2] Calliope, La parole et son traitement automatique. Edition Masson, Paris, 1989.

[3] C.Barras, Reconnaissance de la parole continue : Adaptation au locuteur et contrôle temporel dans les Modèles de Markov Cachés, thèse de doctorat de l'Université ParisVI, Mai 1996.

[4] M.Kunt, Traitement numérique des signaux. Presses polytechniques romandes, Lausanne, 1980.

[5] J.~P.Halton, J.~M.Pierrel, G.Perenou, J.Caelen et J.~L.Gauvain, Reconnaissance automatique de la parole. Edition Dunod, Paris 1991.

[6] R.Boite et M.Kunt, Traitement de la parole. Presses polytechniques romandes, Lausanne, 1987.

[7] F.Coulon, Théorie et traitement des signaux, Presses polytechniques romandes, Lausanne, Edition Georgi, 1984.

[8] Caroline Bousquet-vernhettes, Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique, thèse de doctorat a l'université Paul Sabatier Toulouse III, Septembre 2002.

[9] Salma Jamoussi, Méthodes statistiques pour la compréhension automatique de la parole, thèse de doctorat a l'université Henry Poincaré-Nancy 1, Décembre 2004.

[10] Olivier Le Blouch, Décodage acoustico-phonétique et applications à l'indexation audio automatique, thèse de doctorat a l'université Paul Sabatier Toulouse III, juin 2009.

[11] M.Kunt, M.bellanger, F. Coulon, C.Gueuen, M.Hasler, N.Moreau, M.Vetterli, "Techniques modernes de traitement numérique des signaux", Presses polytechniques et universitaires romandes, Lausanne, 1991.

[12] L.Rabiner, "A tutorial on hidden Markov models and the speech signal", Proceedings of the IEEE, vol.77, no.2, 1989.

[13] L.Rabiner, B-H.Juang, "Fundamentals of speech recognition", Prentice Hall, New Jersey, 1993.

[14] S.Mechhoud, « Identification du locuteur en mode indépendant du texte », projet de fin d'études, département d'électronique, ENP, Juin 2010.

[15] H.Takhdmit , N.Ait Saadi, "Identification du locuteur en mode indépendant du texte", Projet de fin d'études, département d'électronique, ENP, Juin 2005.

[16] Christophe Lévy, Modèles acoustiques compacts pour les systèmes embarqués, thèse de doctorat a l'université d'Avignon et des Pays de Vaucluse, novembre 2006.

[17] Jean Hennebert, Hidden Markov models and artificial neural networks for speech and speaker recognition, these de doctorat a l'école polytechnique fédérale de Lausanne, 1998.

[18] Jeanne Villaneau, « Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée : approche logique pour la compréhension de la parole », thèse de doctorat a l'université Université de Bretagne Sud, décembre 2003.

[19] Georges Linarès, « Reconnaissance automatique de la parole et indexation audiovisuelle », thèse de doctorat a l'université d'Avignon et des Pays de Vaucluse,

[20] G.Blanchet, M.Charbit, "Signaux et images sous Matlab", HERMES science publications, Paris, 2001.

[21] www.claudegabriel.be