UNIVERSITE SAAD DAHLAB DE BLIDA

Faculté des Sciences de l'Ingénieur
Département d'Informatique

MEMOIRE DE MAGISTER

Spécialité : Ingénieries des systèmes et des connaissances

TECHNIQUES D'ANALYSE EN VUE DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Par

HAMADOUCHE Maamar

Devant le jury composé de :

A. GUESSOUM, Professeur, USDB Blida	Président
M. BENAAMANE, Maître de conférence, USDB Blida	Examinateur
F. REGUIGUE, Chargée de cours, USDB Blida	Examinateur
M. GUERTI, Maître de conférence, ENP Alger	Promotrice

Blida, Mai 2008

ملخص

في عملنا هذا, قمنا بعرض بعض تقنيات التحليل الصوتي الأكثر استعمالا في عصرنا خاصة في مجال المعرفة الآلية للكلام , و من بين هته التقنيات نجد تقنية
MFCC (Mel Frequency Cepstrale Coefficients), cepstrale
LPC (Linear Predictive Coding), l'analyse par spectrogrammes la

و تطرقنا أيضا إلى تقنيات المقارنة من اجل تشخيص و معرفة الصوت, ومن بينها مثلا
Les modèles de Markov cachés (HMM pour Hidden Markov Models)
DTW (Dynamic Time Warping) l'alignement temporel dynamique و MMC

أما عملنا فتمحور حول انجاز منظومة للمعرفة الأوتوماتيكية للأرقام العشرة الأولى للغة العربية الكلاسيكية (ARAD (Automatic Recognition of Arabic Digits
النتائج المحصل عليها تقارب 96٪ وهذا بالنسبة للموسوعات الصوتية الثلاث التي سجلناها في محيط عادي

كلمات المفاتيح : تقنيات التحليل الصوتي، المعرفة الأوتوماتيكية للكلام, LPC, MFCC, DTW, HMM , اللغة العربية الفصحى

RESUME

Dans ce travail, Nous présentons les techniques d'analyse du signal vocal les plus répandus actuellement dans le domaine de la Reconnaissance Automatique de la Parole (RAP) telle que la technique des coefficients MFCC (Mel Frequency Cepstrale Coefficients); cepstrale, le calcul par prédiction linéaire LPC (Linear Predictive Coding) et l'analyse par spectrogrammes;.

Nous verrons aussi les méthodes de reconnaissance ou de comparaison qui sont utilisées étroitement avec celles que nous avons citées pour élaborer un Système de RAP (SRAP). Citons à titre d'exemple les modèles de Markov cachés MMC (HMM pour Hidden Markov Models) et l'alignement temporel dynamique DTW (Dynamic Time Warping).

Nous introduisons ensuite notre Système de Reconnaissance Automatique des dix premiers chiffres de l'Arabe Standard ARAD (Automatic Recognition of Arabic Digits). Les résultats obtenus montrent que le système présente un taux de reconnaissance de 96% sur les trois corpus que nous avons enregistrés dans un environnement bruité.

Mots- clés : Techniques d'Analyse, Reconnaissance Automatique de la Parole, LPC, MFCC, DTW, HMM, Arabe Standard.

ABSTRACT

In this memory, We currently present the techniques of analysis of the vocal signal more used in the field of the Automatic speech recognition (RAP) such as the technique of coefficients MFCC (Mel Frequency Cepstrale Coefficients); cepstrale, calculation by linear prediction LPC (Linear Predictive Coding) and analyzes it by spectrograms.

We see also the methods of recognition or comparison which are used narrowly in System of RAP (SRAP). Let us quote as an example the hidden models of Markov MMC (HMM for Hidden Markov Models) and DTW (Dynamic Time Warping).

We introduce then our System of automatically recognition of the first ten digits Arabic Standard ARAD (Automatic Recognition of Arabic Digits). The results obtained show that the system shows a rate of recognition of 96% on the three corpora which we recorded in a disturbed environment.

Words - keys: Techniques of Analysis, Automatic speech recognition, LPC, MFCC, DTW, HMM, Standard Arab.

ABSTRACT

The automatic speech recognition is currently a field in full rise in our daily life to ensure us a new method of communication with our machine, a communication without use of a specific peripheral but only by using the manner most natural to make some, our language natural. The problem of the automatic speech recognition is a subject of topicality and for the moment, only the partial solutions are ready to answer the various tasks which the machine must carry out.

The objective of this work is initially to study the techniques of analysis vocal as well as the method of recognition used in the field of the automatic speech recognition.
And in the second time, we propose a system of recognition automatic of Arab numerals, able to recognize the figures of WAHED with AACHARA as a single-speaker mode, by applying the method of analysis cepstrale (MFCC) in the stage of acoustic analysis and the temporal method of alignment (DTW) to make the recognition.

Key words: parameterization, analyzes word, voice recognition LPC, MFCC, DTW, HMM.

REMERCIEMENTS

Toute ma gratitude et mes vifs remerciements vont à M^{me} M. GUERTI Professeur au département d'Electronique à l'Ecole Nationale Polytechnique ENP Alger, ma Directrice de mémoire, pour sa patience, sa disponibilité, et ses précieux conseils, et qui a su me faire profiter de sa grande expérience, Je souhaite lui transmettre l'expression de ma reconnaissance et ma plus profonde gratitude.

Je remercie tout particulièrement les membres de jury, qui ont accepté de juger mon travail :

- Mr A. GUESSOUM, Professeur à l'université de Blida, Département d'Electronique, d'avoir fait l'honneur de présider mon jury de mémoire;
- M^{elle} F.Z. REGUIEG Chargée de cours au département d'Electronique à l'université de Blida et
- Mr A. BENAAMANE Maître de conférences au département d'Electronique de l'université de Blida.

Je remercie aussi tous mes Enseignants du département d'Informatique de l'Université Saâd Dahlab de Blida.

J'adresse mes sincères remerciements à mes amis de promotion qui ont partagé avec moi les moments de doutes et les moments d'espoir : H. Tebbi, R. Mazari, S. Hassaine, F. Mazari Boufaresse, et Mr M. Hamouda et tous les étudiants de Magister en informatique de l'université de Blida

Je ne pourrais terminer ces remerciements sans me retourner vers les êtres qui me sont les plus chers, qui ont eu un rôle essentiel et continu pendant ma réussite. J'adresse de tout mon coeur mes remerciements à ma chère mère et mon père pour leur soutien moral. Qu'ils trouvent dans ce modeste travail le fruit de leur soutien. Enfin, je voudrais remercier profondément mon frère Salim et mes soeurs, et ma future femme, ainsi que toute la famille.

LISTE DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

TABLE DES MATIERES

INTRODUCTION

La parole est le meilleur moyen de communication entre les humains. Ces dernières années plusieurs recherches ont été effectuées pour réaliser des systèmes automatiques de traitement de la parole.

Le Traitement Automatique de la Parole (TAP) est un domaine de recherche pour lequel un effort important a été consenti au cours des trois dernières décennies. Les problèmes à résoudre sont considérables et de nature fondamentale. Ils sont de plus par essence pluridisciplinaires : traitement du signal, reconnaissance des formes, intelligence artificielle, informatique, phonétique, linguistique, ergonomie, neurosciences qui interviennent à des degrés divers dans les solutions apportées. Ces travaux de longue haleine donnent néanmoins naissance dès à présent à des produits intermédiaires qui trouvent leur place dans des applications pratiques dans le cadre de la communication Homme-Machine. C'est le cas aussi bien en synthèse de la parole (processus de production vocale), qu'en reconnaissance (perception vocale) ou en vérification de locuteur (identification de la personne qui parle). Nous nous intéressons dans notre mémoire plus particulièrement à la Reconnaissance Automatique de la Parole (RAP).

Reconnaître la parole consiste à extraire l'information lexicale contenue dans un signal de parole: on parle de RAP. Le but est de pouvoir utiliser cette information dans une application informatique. On distingue trois grands types de représentation de cette information selon l’application :

- texte (dictée vocale, transcription);
- action (commande vocale, systèmes de dialogue);
- information indexée (annotation, indexation).

Reconnaître la parole se révèle être un problème difficile à cause de sa variabilité, cette dernière est due aux : facteurs intra-locuteurs (co-articulation, variation dans la prononciation, etc.), facteurs inter-locuteurs (physiologie, âge, sexe, émotions, etc.) et à l'environnement (bruit, micro, canal de transmission, présence d'autres locuteurs, etc.).Intuitivement, il est évident que *parler* et *entendre* sont très liés, puisque l'action d'entendre ne se résume pas à déchiffrer des caractères mais plutôt à donner du sens à ce qui a été émis.
Du point de vue informatique, les difficultés de la RAP sont liées bien évidemment à la variabilité des locuteurs, mais aussi au fait qu'on ne sait pas encore très bien modéliser l'énorme masse de connaissances et d'informations utiles à la reconnaissance.

Un Système de RAP (SRAP) est un logiciel dont l'entrée est un signal acoustique et qui fournit en sortie une représentation symbolique de ce signale, ou bien une étiquette. L'objectif vers lequel tendent les SRAP est de donner du sens à la parole. Différents SRAPs ont été développés ces dernières années, couvrant des domaines aussi vastes que la reconnaissance de mots clés sur lignes téléphoniques, les systèmes de dictée vocale, les systèmes de commande et contrôle sur PC, en allant jusqu'aux systèmes de compréhension du langage naturel (pour des applications limitées).

Pour présenter notre mémoire, nous procédons comme suit :

- au chapitre1, nous présentons quelques caractéristiques de la parole et nous abordons la problématique d'analyse qui est essentielle pour notre application;
- le chapitre 2 présente les théories de base dans le domaine de la RAP ainsi que les différentes techniques utilisées;
- le chapitre 3 introduit les techniques d'analyse en vue de la RAP les plus répandues actuellement.
- le chapitre 4 décrit le SRAP que nous avons développé ainsi que les résultats obtenus;

Finalement, nous décrivons quelques extensions possibles de notre système, nous présentons aussi les résultats que nous avons obtenus pour la validation de notre système, et nous conclurons notre travail en résumant les idées principales du mémoire et présentant quelques axes à explorer dans les recherches futures.

CHAPITRE 1
GENERALITES SUR LA PAROLE

1.1. Introduction

La parole est un moyen essentiel pour la communication entre les humains, c'est la capacité de communiquer la pensée par un système de sons articulés émis par les organes de la phonation. Ces dernières années et avec l'avancée spectaculaire des moyens de communication et des outils informatiques, nous tentons une communication Homme-Machine moyennant ce même outil privilégié à l'humain. Dans ce chapitre nous verrons la parole depuis sa production jusqu'à sa perception, nous détaillerons le signal vocal en abordant ses caractéristiques pertinentes et nous ferons un aperçu sur la langue arabe en abordant ses propres particularités.

1.2. La parole

La parole peut être vue comme une suite de sons produits soit par des vibrations des cordes vocales (source quasi périodique de voisement), soit par une turbulence créée par l'air s'écoulant dans le conduit vocal, lors du relâchement d'une occlusion ou d'une forte constriction de ce conduit (sources de bruit non voisées) [1].

La parole correspond à une variation de la pression de l'air causée par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié qui de nos jours est le plus souvent numérisé [2].

La parole est formée de phonèmes et de transitions entre ces phonèmes. Plusieurs types de phonèmes existent : les voyelles, les consonnes fricatives, les consonnes plosives, les nasales et les liquides. Les transitions acoustiques correspondent à des transitions dans l'appareil de

production de l'état correspondant au premier phonème à l'état correspondant au suivant [3].

1.2.1. L'échantillonnage

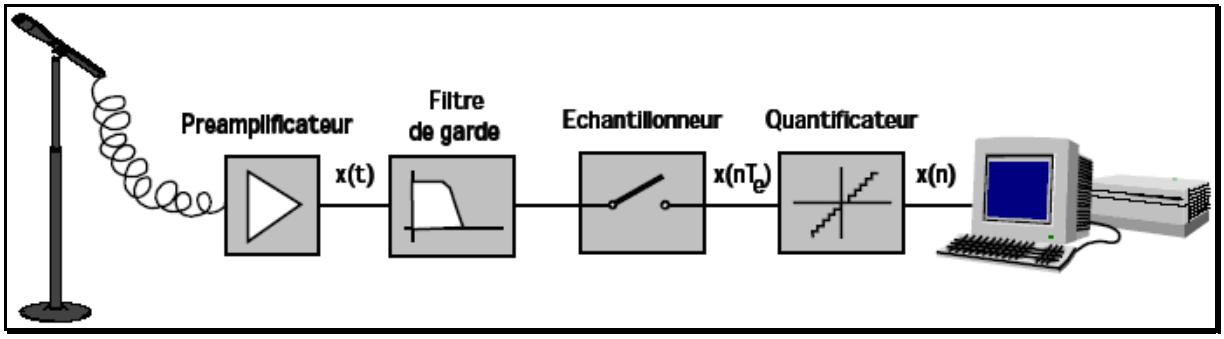


Figure 1.1 : Enregistrement numérique d'un signal acoustique.

L'échantillonnage transforme le signal à temps continu x (t) en signal à temps discret x (n) défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage; celle-ci est elle-même l'inverse de la fréquence d'échantillonnage. Pour ce qui concerne le signal vocal, le choix de cette fréquence d'échantillonnage résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz. Il faut donc en principe choisir une fréquence égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon. Cependant, le coût d'un traitement numérique, filtrage, transmission, ou simple enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable (Figure1.1).

Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit une fréquence d'échantillonnage égale à 8 kHz. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 6 à 16 kHz. Par contre, pour le signal audio (parole et musique), on exige une bonne représentation du signal jusqu'à 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz [2].

1.2.2. Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un spectrogramme. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre

de pondération (Figure.1.2). Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms); ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales.

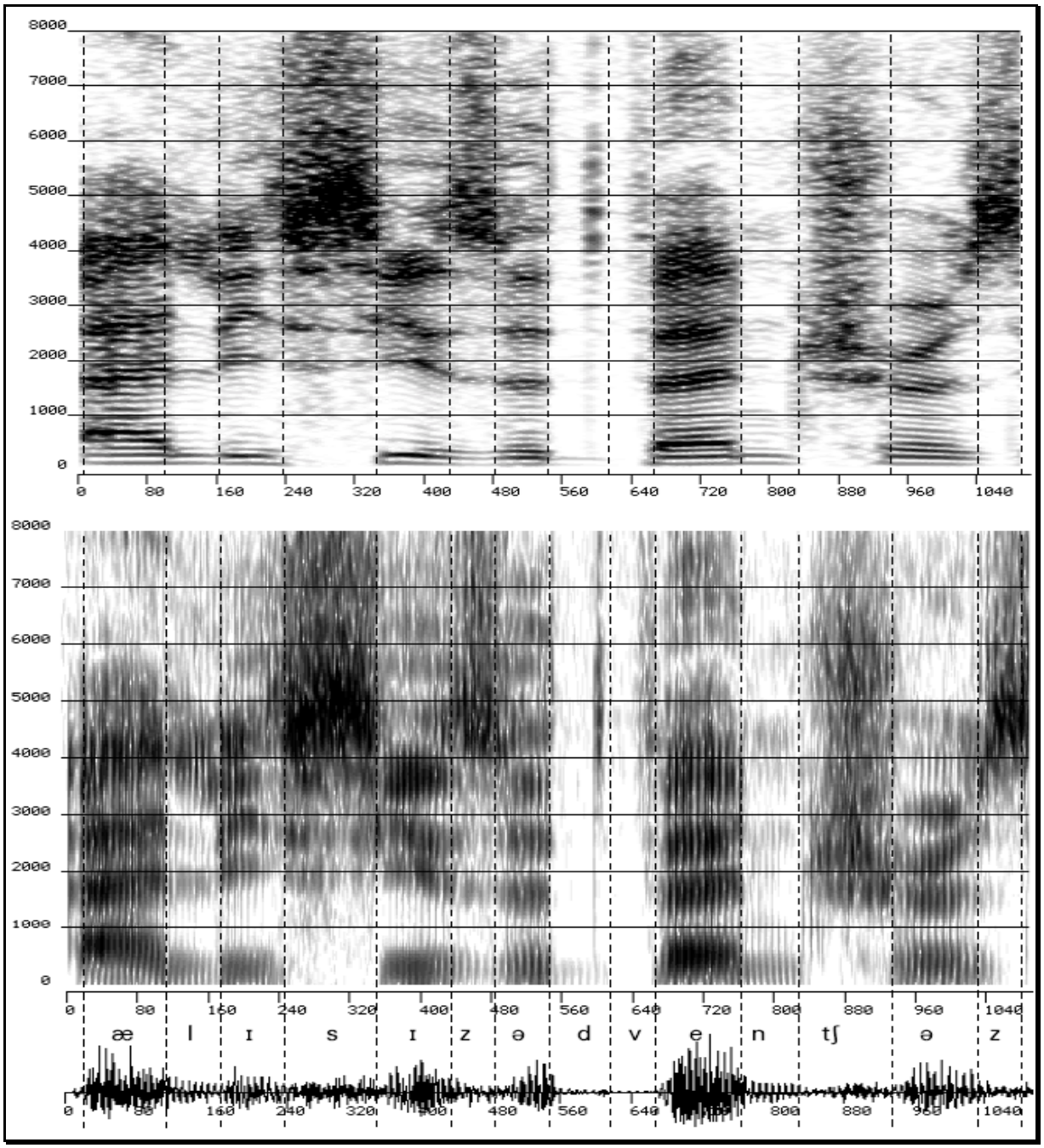


Figure 1.2 : Spectrogrammes à large bande (en bas) et à bande étroite (en haut) [4].

1.3. Comment se produit la parole ?

La parole est un signal réel (4D), continu, d'énergie finie et non stationnaire. Sa structure est complexe et variable dans le temps [5].

Pour comprendre le mécanisme de production de la parole il faut voir de pré l'appareil phonatoire humain en allant depuis les poumons jusqu'aux lèvres, puisque "La production de la parole n'est pas réalisée par un système

propre, mais est assurée conjointement par les organes de la respiration et de la déglutition” [6].

A l’instar des instruments à vent dont l’embouchure est munie d’une anche, l’appareil de la phonation se décompose en trois parties essentielles. La première est la source énergétique de la voix : c’est la "soufflerie", autrement dit l’appareil respiratoire dans son ensemble. La deuxième correspond (en première approche) à la anche des instruments à vent, il s’agit du larynx, à l’origine de la vibration sonore, et la troisième, enfin, est formée par les diverses cavités de résonance dans lesquelles le son émis au niveau du larynx va se voir amplifié modifié (Figue 1.3)[7].

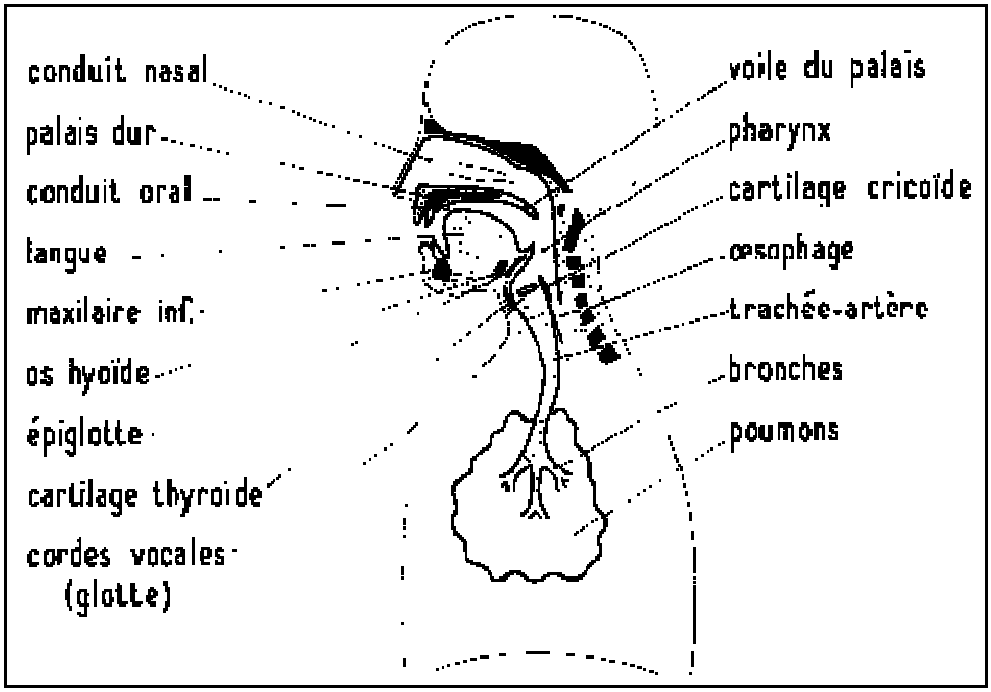


Figure 1.3 : l'anatomie de l'appareil phonatoire humain

Lorsque l’on parle ou que l’on chante, le mouvement respiratoire s’adapte d’une manière très différente à la conformation de la respiration normale. Alors que la phase expiratoire, où l’air est expulsé des poumons, se rallonge considérablement, la durée de la phase inspiratoire est, elle, raccourcie. En corollaire, les volumes d’air mobilisés dans les poumons sont beaucoup plus importants que ceux observés dans le cas de la respiration. De même, la pression de l’air expulsé des poumons est plus grande, car le

flux expiré rencontre avant sa sortie l'obstacle des cordes vocales, dont il va provoquer la vibration [7].

On peut ainsi découpé l'appareil phonatoire en trois parties (figure 1.4), où chacune d'elle renferme plusieurs organes participant à la réalisation de l'une des trois source citées ci avant, qui sont [5] :

- *Partie sub-glottique* (source d'énergie): elle contient les poumons et la trachée artère, sa fonction est de fournir l'énergie nécessaire à la phonation en insufflant l'air vers la partie glottique ;

- *Partie glottique ou larynx* (source d'excitation): ensemble de cartilages, ligaments et muscles qui englobe 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, en plus elles libèrent un flux d'air vers la partie supraglottique ;

- *Partie supraglottique ou le conduit vocal* (source de résonance) contenant :

  - Cavités orales (pharyngienne et buccale) à géométrie variable en fonction des articulateurs: langue, mâchoire inférieure et lèvres.
  - Cavités nasales à géométrie fixe peuvent être couplées aux cavités orales par abaissement du voile du palais.

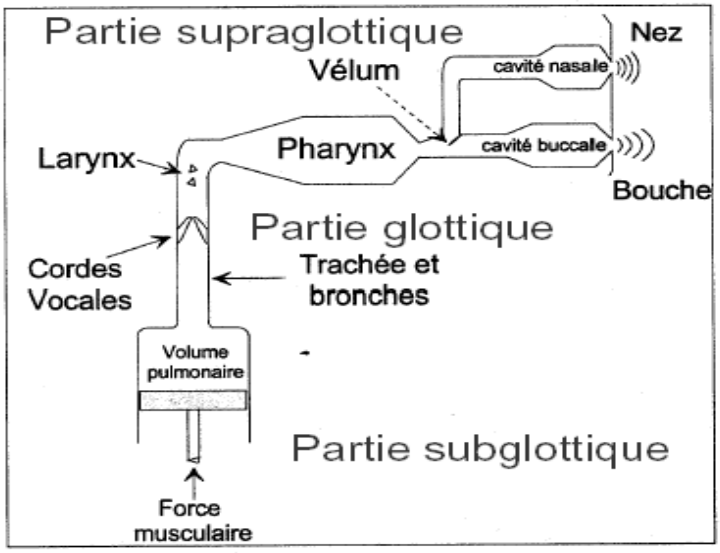


Figure 1.4 : representation schématique de la production de la parole [7]

Donc on peut dire que la production de la parole est le résultat de deux fonctions mécaniques de base: La production d'un signal acoustique, par une ou deux sources acoustiques, ou *phonation*, et sa modulation par les cavités supraglottiques, orales et/ou nasales, ou *articulation* [6].

1.3.1. La phonation

Le processus de la phonation comporte deux étapes essentielles: La génération d'une énergie ventilatoire, qui va être utilisée pour mettre en mouvement les cordes vocales, et leur vibration, qui donne naissance à tous les sons voisés, et/ou pour générer des bruits.

La phonation est réalisée sur le temps expiratoire de la respiration; dès la fin de l'inspiration, il ne se produit non pas un relâchement de la cage thoracique, comme dans l'expiration normale, mais une mise en tension solide du thorax par le jeu antagoniste des muscles inspirateurs et expirateurs pour aboutir à la fin de la phrase à une posture expiratoire [6].

1.3.2. L'articulation

La réalisation d'une gestuelle articulatoire au niveau du conduit vocal et des fosses nasales est le deuxième processus essentiel à la production de la parole. Cette modulation du signal acoustique est réalisée par le mouvement du conduit vocal et des fosses nasales, assurée par les articulateurs, structures anatomiques et fonctionnelles comprenant des muscles, des cartilages et des os [6].

1.3.3. Les résonateurs et les articulateurs

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, la cavité buccale, la cavité labiale et les fosses nasales. La présence ou l'absence d'articulateurs sur le parcours de la colonne d'air modifie la nature du son produit. C'est, entre autres, en classant ces obstacles éventuels que la phonétique articulatoire dégage les différentes classes de sons. Pour un petit

nombre de réalisations, comme les clics, l'air ne provient pas des poumons, mais de l'extérieur, par inspiration. La figure 1.5 montre les résonateurs et les articulateurs de l'appareil phonatoire humain.

On peut dire en résumé que l'air contenu dans les poumons représente la source d'énergie utilisée pour produire les sons et, ce flux d'air sous pression parvient à travers la trachée artère jusqu'au conduit vocal, aux fosses nasales, aux organes d'articulation (langue, lèvres...) qui vont avoir chacun leur rôle dans la production de la parole.

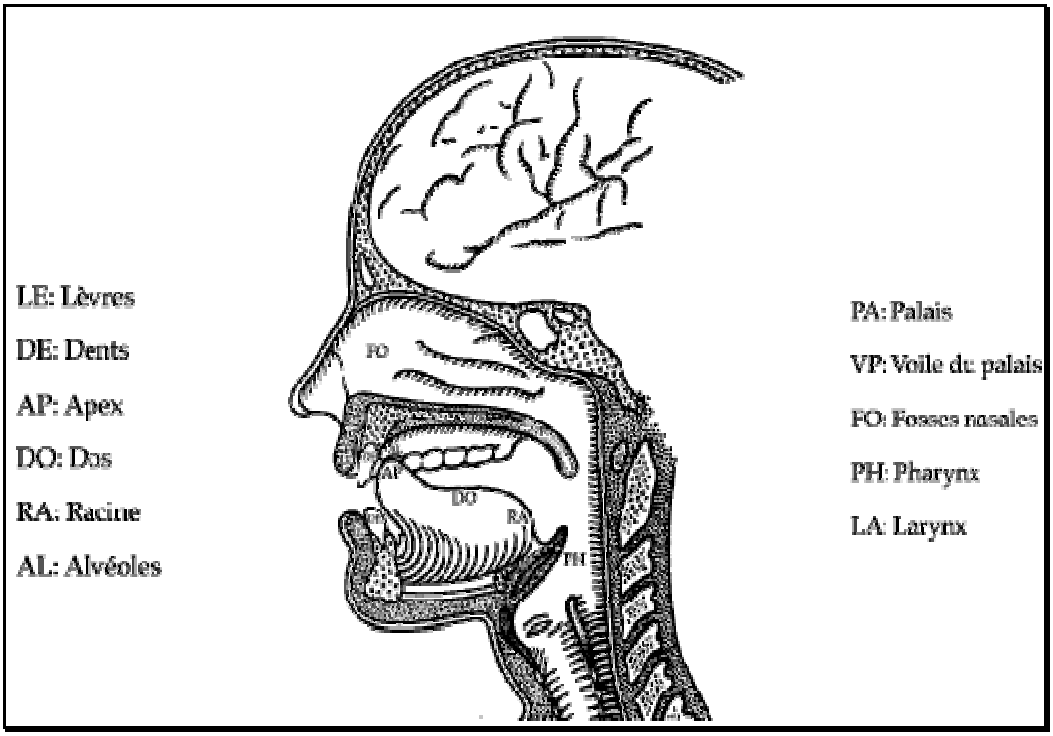


Figure 1.5 : Résonateurs et articulateurs [6]

La bouche est un autre élément qui participe à la production d'un son grâce à la mâchoire supérieure formée par les alvéoles (la partie interne de la gencive), le palais dur (au milieu) et le voile du palais. Au niveau de la partie inférieure, la mobilité de la mâchoire inférieure et la langue effectuent le déplacement. Pour la langue, on distingue la pointe de la langue et le dos de la langue. Les lèvres et les dents participent aussi à la phonation [39]. La parole sera donc très dépendante des caractéristiques physiques de la personne qui parle : âge, taille, sexe, etc. Ainsi [8] :

- la phonation désigne le mécanisme de production des sons du langage;

- un processus expiatoire permet d'amener l'air jusqu'au larynx;

- la vibration des cordes vocales et/ou le frottement de l'air dans le conduit vocal sont responsables des sons émis;

- le conduit vocal filtre les sons émis au niveau du larynx;

- les articulateurs, par la déformation de la forme du conduit vocal, permettent de moduler les sons produits.

1.3.4. Les types de sons produits

La vibration quasi-périodique des cordes vocales permet la production de toutes les voyelles et aussi de certaines consonnes dites sonores ou voisées comme les sons [b] et [d] [39]. Les différents états de l'appareil phonatoire humain déterminent les différents types (nature) des sons produits (figure 1.6) :

- Les voyelles sont le résultat de l'ouverture du conduit vocal, les cordes vocales vibrent (sons voisés) et la forme des cavités (essentiellement la bouche) modifie le timbre. Les voyelles sont dites orales ou nasales selon que la cavité nasale n'est pas ou est mise en parallèle à la cavité buccale.

- Les consonnes se caractérisent par un rétrécissement de l'appareil phonatoire, les cordes vocales peuvent vibrer ou laisser passer librement l'air (sons voisés et non voisés), Les consonnes sont fricatives si le rétrécissement est partiel ou occlusives (plosives) si une occlusion totale apparaît dans l'appareil phonatoire, causant une augmentation de la pression et un relâchement brutal de celle-ci lors de l'ouverture.

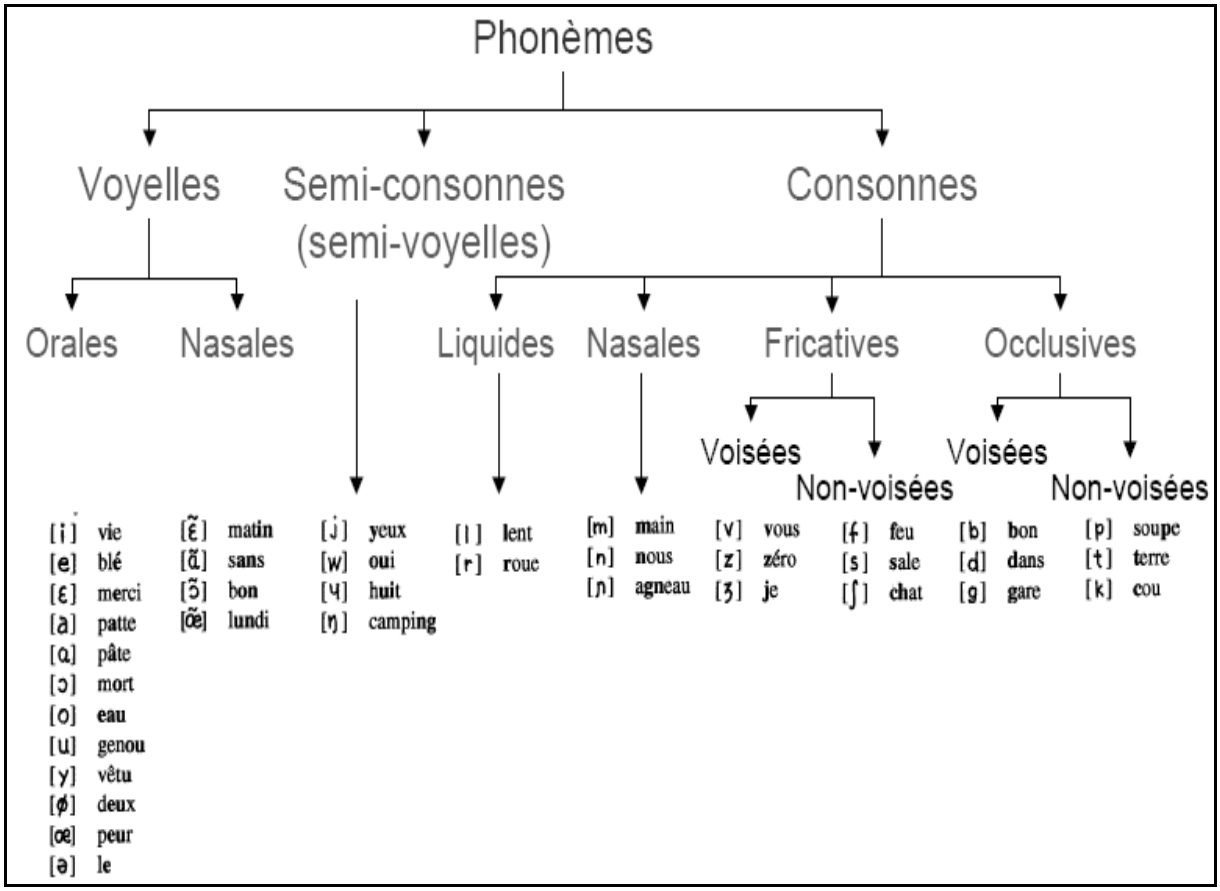


Figure 1.6 : Phonèmes de la langue française [5].

1.3.5. Perception humaine

L'être humain qualifie les sons suivants différents facteurs liés à des mesures physiques [7] :

- Intensité : L'intensité du son est liée à la pression de l'air en amont du larynx ;
    - cause: une énergie plus ou moins forte en provenance du diaphragme ;
    - conséquence : cette énergie provoque une pression plus ou moins forte de l'air sous la glotte ;
    - résultat sonore : la variation de l'amplitude du son émis est plus ou moins grande.

- hauteur : qui est la qualité d'un son plus ou moins grave ou aigu dépend du nombre de vibrations ;

- cause : périodicité plus ou moins grande du battement des cordes vocales;
- conséquence : cette périodicité produit une fréquence de la variation de pression de l'air plus ou moins grande ;
- résultat sonore: le son est grave ou aigu.

- timbre : qui est la qualité d'un son plus ou moins grave ou aigu dépend du nombre de vibrations ;

- cause : positionnement du voile du palais ;
- conséquence : le son est plus ou moins riche en harmoniques graves ou aiguës ;
- résultat sonore : le son semble clair ou pas, "riche", "sombre" ou "couvert".

1.3.6. Le système auditif

L'appareil auditif humain comprend l'oreille externe, l'oreille moyenne, et l'oreille interne (Figure1.7). Le conduit auditif relie le pavillon au tympan; c'est un tube acoustique de section uniforme fermé à une extrémité, son premier mode de résonance est situé vers 3000Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences. Le mécanisme de l'oreille moyenne (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 impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche 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 [4].

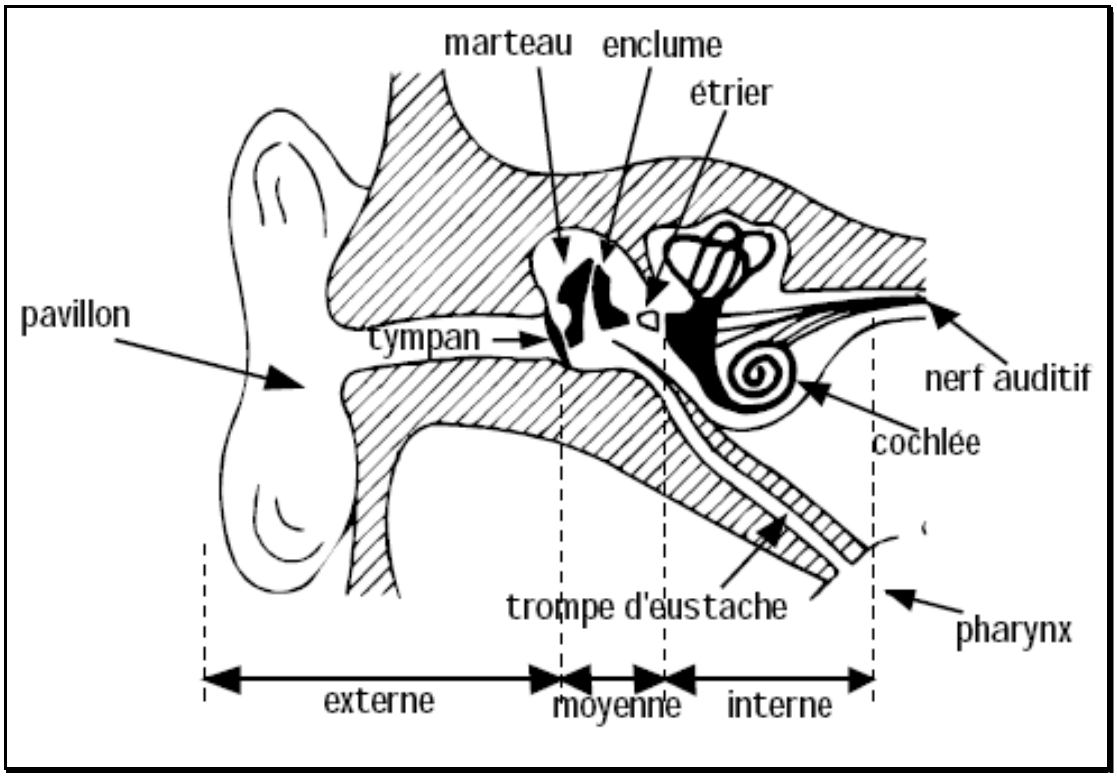


Figure 1.7 : le système auditif humain [9]

Donc les vibrations mécaniques du signal sont converties en impulsions nerveuses du nerf auditif par les cellules ciliées au niveau de la cochlée [9].

Il est à noter que l'oreille humaine ne répond pas de la même manière à toutes les fréquences. La figure1.8 présente le champ auditif humain, situé entre deux courbes; la courbe du seuil de l'audition et celle du seuil de la douleur. La limite supérieure en fréquence est d'environ 16-20kHz [9] (variable selon les individus) se qui donne une fréquence d'échantillonnage maximale, selon le théorème de Shannon, pour un signal auditif d'environ 40kHz.

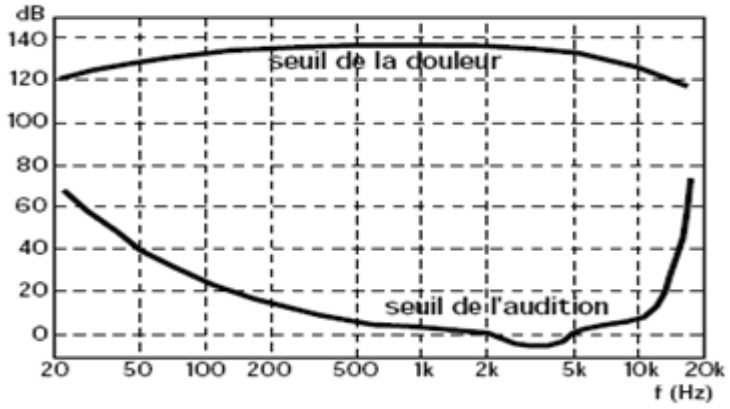


Figure 1.8 : Champ auditif humain [4]

1.4. Caractéristiques pertinentes du signal de la parole

Pour en revenir sur les caractéristiques principales du signal vocale il faut bien comprendre la nature de la parole en temps qu'un phénomène vibratoire produit par un système articulatoire de l'appareil phonatoire humain. On peut résumé les principales caractéristiques du signal vocale dans ce que suit :

- *la non stationnarité* : Le signal vocal n'est pas un signal stationnaire puisque le conduit vocal se déforme d'une façon continue et les paramètres du modèle sont donc variables dans le temps. Toutefois, les déformations sont suffisamment lentes pour que les coefficients de la fonction de transfert puissent être modélisés comme quasi-constants pendant des intervalles de temps de l'ordre de *10 ms* [5];

- *la continuité* : Le langage oral est une suite continue de sons sans séparation entre les mots. Les silences correspondent en général à des pauses de respiration dont l'occurrence est aléatoire. Il peut très bien y avoir des intervalles de silences au milieu d'un mot et aucun intervalle entre deux mots successifs. Par conséquent, il est très difficile de déterminer le début et la fin des mots composant la phrase [1];

- *la variabilité* : Le signale de parole présente une variabilité lui permet de différencier entre les locuteurs, ce qu'on appel la variabilité d'une

personne à une autre ou bien la variabilité inter-locuteurs. Viens s'ajouter à cette dernière, qui est utile pour différencier les locuteurs, autres types de variabilité : variabilité intra-locuteurs, variabilité due aux conditions d'enregistrement et de transmission du signal de parole (bruit ambiant, microphone utilisé, lignes de transmission) et variabilité due au contenu linguistique ;

- Variabilités inter-locuteurs proviennent des différences physiologiques (différences dimensionnelles du conduit vocal, fréquence d'oscillation des cordes vocales) et de différences de style de prononciation (p.ex. accent, niveau social). Certaines de ces différences qui influencent la représentation de chaque locuteur, nous permettent de les séparer [10];
- Variabilités intra-locuteurs font que la voix dépend de l'état physique et émotionnel d'un individu. La voix humaine varie avec le temps ou les conditions physiologiques et psychologiques du locuteur. Cependant, ces variations intra-locuteurs ne sont pas identiques pour tous les humains. En effet, hormis les variations lentes de la voix dues au vieillissement, certains phénomènes extérieurs tels que l'état de santé d'une personne ont une influence variable sur sa voix [10];
- Il y a un autre type de variabilité qu'on appel la variabilité contextuelle qui est due au phénomène de coarticulation des sons entre eux tels que deux sons voisins peuvent influencer mutuellement ; cette variabilité est appelée aussi *la variabilité due à l'environnement* et cela puisque l'environnement peut diminuer le signal vocal généré sans que le locuteur modifier son mode d'élocution.

1.5. La prosodie

Le terme "prosodie" désigne collectivement l'intonation (mesurée par la fréquence fondamentale), le rythme ou la durée, et l'intensité du signal de

parole, dont les valeurs évoluent au cours du temps d'une manière propre à la langue et au locuteur.

La prosodie est une branche de la linguistique de l'expression (phonétique et phonologie) qui s'attache à la description et à la représentation physique et formelle des éléments phoniques systématiques du langage différents des phonèmes tels que l'accent, l'intonation, dont la manifestation concrète est associée à des variations de la fréquence fondamentale F_0, de la durée et de l'intensité (paramètres prosodiques physiques) qui sont perçues comme des changements de hauteur ou de mélodie, de longueur et de volume sonore (paramètres prosodiques subjectifs) [11].

1.5.1. La fréquence fondamentale F_0 ou pitch

La Fréquence Fondamentale est la fréquence de vibration des cordes vocales, elle varie d'une personne à une autre en fonction de la longueur et de la tension des cordes vocales de chaque personne. Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale ou pitch. La figure 1.9 donne l'évolution temporelle de la fréquence fondamentale de la phrase *"les techniques de traitement de la parole"*.

On constate qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps. Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants [4], tandis que les sons non voisés sont associés à une fréquence fondamentale nulle [2].

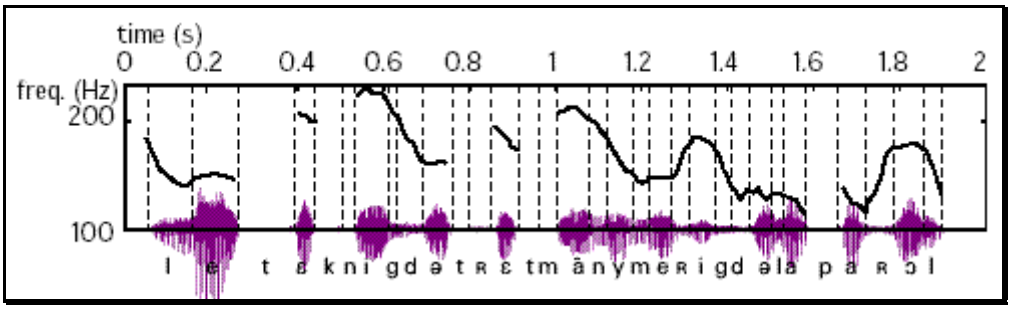


Figure1.9 : Evolution de la fréquence de vibrations des cordes vocales dans la phrase "les techniques de traitement numérique de la parole".

1.5.2. La hauteur

La hauteur qui est la qualité d'un son plus ou moins grave ou aigu dépend du nombre de vibrations. En effet, il faut savoir qu'en matière d'acoustique un corps ne peut émettre un son que s'il vibre. Ces vibrations sont alors transmises à l'oreille par l'intermédiaire de l'air et leur fréquence peut varier entre 16 et 20000 vibrations par seconde (l'unité de fréquence est le Hertz, symbole : Hz). Ainsi plus le son est aigu, plus les vibrations sont nombreuses et inversement [12]. La hauteur du son perçu est dite "grave" entre 15 et 200 Hz, "medium" entre 200 et 2.000 Hz, "aiguë" au delà de 2.000 Hz, ces valeurs caractérisant d'ailleurs plutôt des zones que des frontières.

1.5.3. La durée (rythme)

La durée ou la tenue des sons dépend de la pression de l'air expiré. Généralement la durée d'une unité est mesurée par le nombre des trames (phonèmes) qu'elle contient, et pour calculer la durée de chaque trame il faut fixer deux événements sur le signal de parole qui délimitent les repères initial et final de cette trame.

1.5.4. L'intensité (énergie)

L'intensité fait distinguer un son fort d'un son faible et est liée avec l'amplitude des vibrations. Les acousticiens l'expriment en décibels, notés dB. Le Bel est une unité d'usage, exclusivement scientifique, exprimant la valeur du logarithme décimal du rapport de deux grandeurs physiques. Son dixième,

le décibel (noté dB) est entré aujourd'hui dans le langage courant pour y caractériser l'unité de bruit.

Voici à titre d'exemple l'intensité de quelques sources sonores : un bruissement de feuilles: de 0 à 10 dB; une conversation normale: 20 à 50 dB; une discussion animée: 50 à 65 dB; le passage d'un train: 65 à 90 dB; le bruit du tonnerre: 90 à 110 dB et d'un avion à réaction au décollage: 110 à 140 dB. Un bruit devient douloureux à l'oreille à partir de 120 dB et au dessus de 160 dB on peut parler de sons destructeurs! [12]

1.6. La classification des sons du langage

L'air nécessaire pour la production des sons sort des poumons et passe par la trachée, en haut de la trachée se trouve une boîte en cartilage qu'on appelle le larynx. Suspendues dans le larynx on trouve deux bandes de tissu élastiques, qu'on appelle les cordes vocales ou la glotte. Si les cordes vocales sont ouvertes, on entend un son non voisé ou sourd comme *[p]*. Si elles se rapprochent et vibrent, on a un son voisé comme *[v]*.

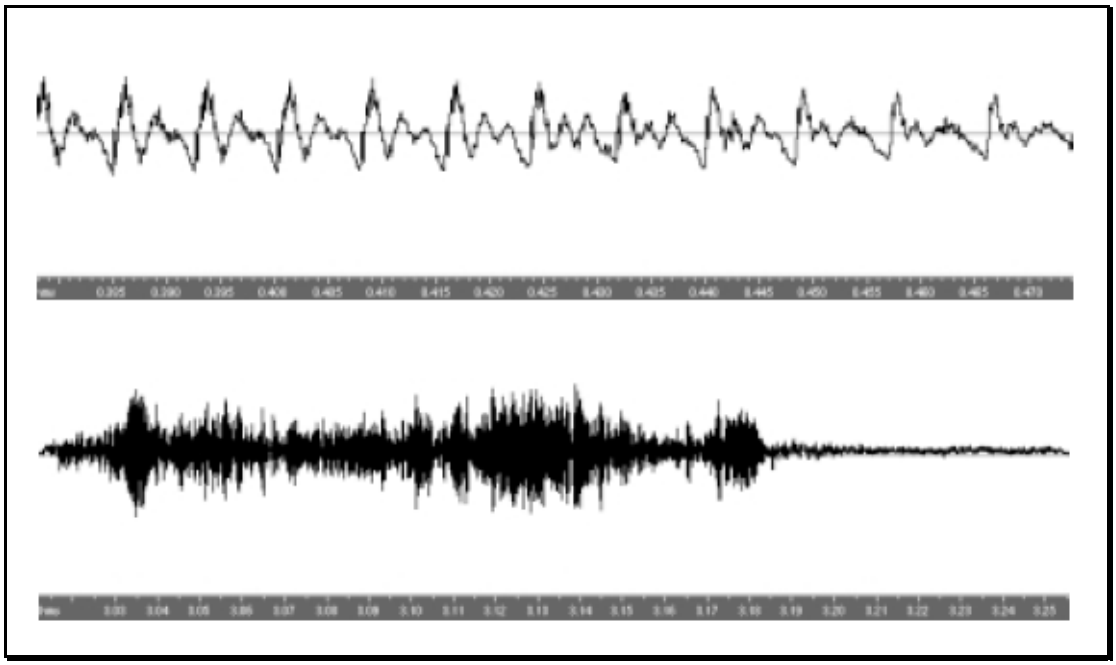


Figure 1.10 : Exemples de son voisé (haut) et non-voisé (bas) [2].

1.6.1 Les sons voisés

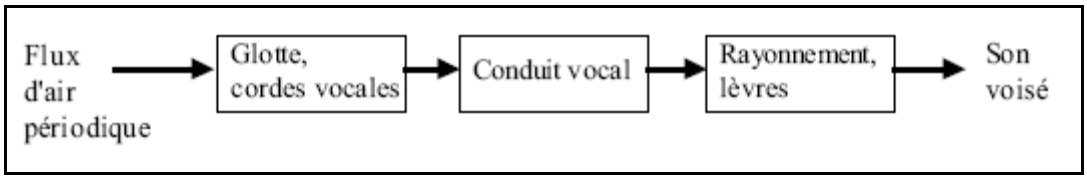


Figure 1.11 : Processus de production de la parole dans le cas des phonèmes voisés.

Dans le cas des phonèmes voisés le flux d'air *p* est un train d'impulsion de période *N*. Ce flux d'air est modifié par les contributions glottales *g*, rayonnement (lèvres) *r* et celle du conduit vocal *v* (figure 1.11). Le signal de parole *y* résultant est la convolution de *p* par les réponses impulsionnelles *g*, *r*, *v* des trois parties du processus de production de la parole [13] comme le présente l'équation suivante :

$$Y = P * g * v * r \qquad (1.1)$$

1.6.2 Les sons non voisés

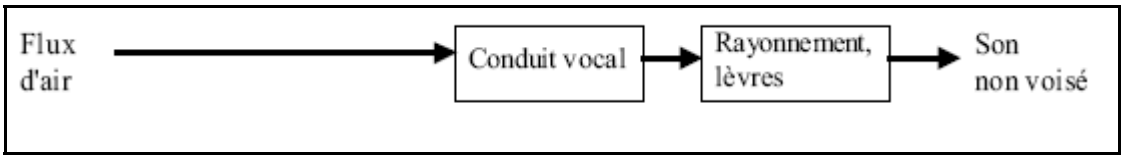


Figure 1.12 : Processus de production de la parole dans le cas des phonèmes non voisés.

Le son non voisé peut être considéré comme un bruit blanc qui résulte d'un écoulement turbulent de l'air à travers le conduit vocal (figure 1.12). Sa forme d'onde ne présente aucune périodicité. Dans ce cas les cordes vocales ne vibrent pas. Le flux d'air u est considéré comme un bruit blanc [13] :

$$Y = u * v * r \qquad (1.2)$$

1.7. Modèle de la production de la parole

Comme pour tous les instruments à vent les systèmes de description de la voix se basent sur le modèle de production acoustique source-filtre [14].

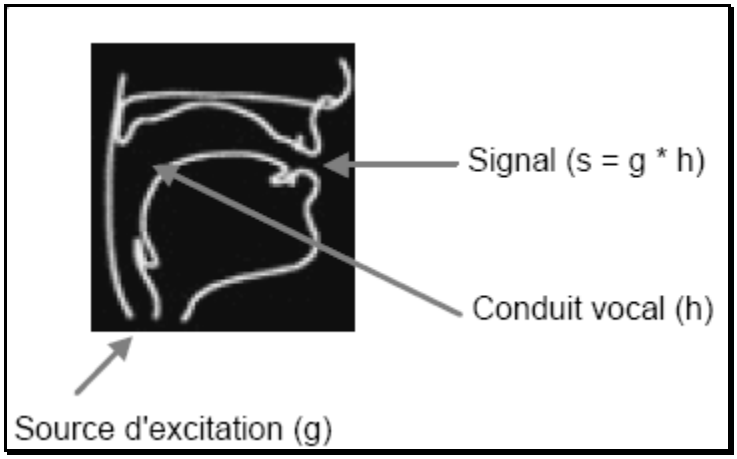


Figure 1.13 : Filtre linéaire: produit de convolution (s = g * h) [15].

Le signal de parole est provoqué par des mécanismes complexes issus de plusieurs sources, il peut être modélisé soit par une source d'impulsions périodiques (pour la production des voyelles), soit par une source de bruit blanc (pour la production des consonnes) [15]. Le conduit vocal, longueur de 17 cm en moyenne chez l'homme adulte [14] peut être modélisé par un filtre qui varie en fonction du temps. Ce filtre est composé de plusieurs résonances (formants). La production de la voix peut donc être représentée par un modèle source-filtre correspondant au signal d'excitation et au conduit vocal (voir figure 1.14). Dès lors, la structure fine du spectre est contrôlée par la source et l'enveloppe spectrale est contrôlée par le filtre. Celui-ci est appelé filtre formantique et comporte généralement deux résonances : les deux premiers formants suffisent à caractériser une voyelle [16].

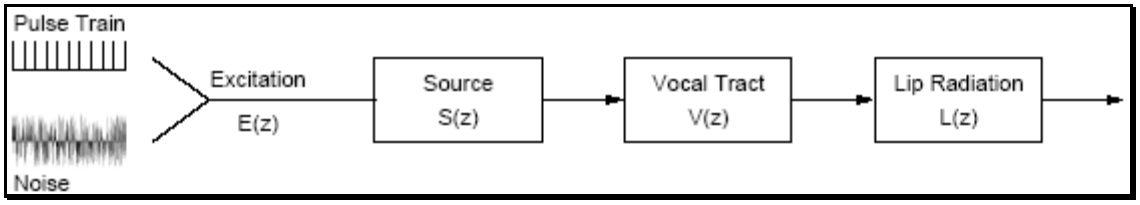


Figure 1.14 : Modele de production de la parole.

La source donne donc au son hauteur ou bruit et puissance, tandis que le conduit vocal agit comme un filtre acoustique. L'information phonétique est contenue dans le type de source utilisé et dans l'enveloppe spectrale du filtre. Il faut donc séparer les contributions de chacun dans le signal pour la décrire efficacement [14].

1.8. Vue d'ensemble de la langue arabe

La langue arabe est une langue sémitique, elle est parmi les langues les plus anciennes dans le monde [17]. Par ses propriétés morphologiques et syntaxiques la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique de la langue, l'arabe doit sa formidable expansion à partir du 7eme siècle grâce à la propagation de l'islam et la diffusion du Coran, les recherches pour le traitement automatique de l'arabe ont débuté vers les années 1970 [18]. Les premiers travaux concernaient notamment les lexiques et la morphologie.

L'arabe classique standard a 34 phonèmes parmi lesquels 6 sont voyelles et 28 sont des consonnes [19]. Les phonèmes arabe se distinguent par la présence de deux classes qui sont appelées pharyngales et emphatiques. Ces deux classes sont caractéristiques des langues sémitiques comme l'hébreu [20]. Les syllabes permises dans la langue arabe sont : *[CV]*, *[CVC]* et *[CVCC]*. Où le [V] désigne voyelle courte ou longue et le [C] représente une consonne [19]. La langue arabe comporte cinq types de syllabes classées selon les traits ouvert/fermé et court/long. Une syllabe est dite ouverte (respectivement fermée) si elle se termine par une voyelle (respectivement une consonne). Toutes les syllabes commencent par une consonne suivie d'une voyelle et elles comportent une seule voyelle. La syllabe *[CV]* peut se trouver au début, au milieu ou à la fin du mot [21].

1.9. Spécificité de la langue arabe

Comme précédemment évoqué, l'alphabet de la langue arabe comporte 28 consonnes (Tableau 1.1). La langue arabe s'écrit de droite à gauche, les lettres peuvent changés leur forme de présentation selon leur position (au début, au milieu ou à la fin du mot). Le Tableau 1.2 montre les variations de la lettre ع (Ayn). Toutes les lettres se lient entre elles sauf (د ذ, ز, ر, و, ا,) qui ne se joignent pas à gauche .

Transcription phonétique		Entendre la lettre	Lettre arabe
ḍād	ḍ	dâd	ض
ṭāʾ	ṭ	ta	ط
ẓāʾ	ẓ	tha	ظ
ʿaïn	ʿ	'ain	ع
ġaïn	ġ	ghain	غ
fāʾ	f	fâ	ف
qāf	q	qâf	ق
kāf	k	kâf	ك
lām	l	lâm	ل
mīm	m	mîm	م
nūn	n	nûn	ن
hāʾ	h	ha	ه
wāw	w	wâw	و
yāʾ	y	yâ	ي

Transcription phonétique		Entendre la lettre	Lettre arabe
		alif	ا
alif	a	ba	ب
bāʾ	b	ta	ت
tāʾ	t	ta	ث
ṯāʾ	ṯ	jim	ج
jīm	j	ha	ح
ḥāʾ	ḥ	ha	خ
ḫāʾ	ḫ	dal	د
dāl	d	dhal	ذ
ḏāl	ḏ	r	ر
rāʾ	r	zay	ز
zīn/zāy	z	sîn	س
sīn	s	shîn	ش
šīn, shin	š, sh	sâd	ص
ṣād	ṣ		

Table 1.1 : les 28 lettres de la langue arabe

à la fin d'une lettre non joignable	à la fin	au milieu	au début
ع	ع	ـعـ	عـ

Table 1.2 : Exemple de variations de la lettre ع Ayn[18]

lettre	nom	fin	milieu	début	phonétique
ا	alif	ـا	ـا	ا	a:
ب	ba	ـب	ـبـ	بـ	b
ت	ta	ـت	ـتـ	تـ	t
ث	tha	ـث	ـثـ	ثـ	θ
ج	jim	ـج	ـجـ	جـ	ʤ , ʒ , g
ح	ha	ـح	ـحـ	حـ	ħ
خ	Ha	ـخ	ـخـ	خـ	x
د	dal	ـد	ـد	د	d
ذ	dhal	ـذ	ـذ	ذ	ð
ر	ra	ـر	ـر	ر	r
ز	za	ـز	ـز	ز	z
س	sin	ـس	ـسـ	سـ	s
ش	shin	ـش	ـشـ	شـ	ʃ
ص	Sad	ـص	ـصـ	صـ	sˤ
ط	Ta	ـط	ـطـ	طـ	tˤ
ظ	Za	ـظ	ـظـ	ظـ	zˤ , ðˤ
ع	ayn'	ـع	ـعـ	عـ	ʔˤ
غ	ġayn	ـغ	ـغـ	غـ	ɣ
ف	fa	ـف	ـفـ	فـ	f
ق	qaf	ـق	ـقـ	قـ	q
ك	kaf	ـك	ـكـ	كـ	k
ل	lam	ـل	ـلـ	لـ	l
م	mim	ـم	ـمـ	مـ	m
ن	nun	ـن	ـنـ	نـ	n
ه	ha	ـه	ـهـ	هـ	h
و	waw	ـو	ـو	و	w , u:
ي	ya	ـي	ـيـ	يـ	j, i:
ء	hamza		أ ؤ إ ئ		ʔ

Table 1.3 : variation des lettres de la langue arabe.

Concernant la prononciation, Les lettres transcrites avec une majuscule dans le Tableau1.3 sont emphatiques (elles sont aussi transcrites avec un point sous la lettre) : elles se prononcent comme si l'on avait la bouche pleine *Ha* se prononce comme la *jota* espagnole (ou le *ch* allemand). Les lettres transcrites *tha*, *dhal* se prononcent comme le *th* anglais : (elles sont aussi transcrites avec un tiret sous la lettre), de sorte que *tha* se prononce comme le *th* de *thing* et *dhal* comme le *th* anglais de *this*.

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres (ـَ, ـِ, ـُ, ـْ)
Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier des mots ayant la même représentation. Le Tableau 1.4 donne un exemple pour les mots كتب et مدرسة. Cependant, les voyelles ne sont utilisées que pour des textes sacrés et didactiques. Les textes courants rencontrés dans les journaux et les livres n'en ne comportent habituellement pas. De plus certaines lettres comme ا Alef peuvent symboliser le آ ,أ ou إ; de même que pour les lettres ى et ه qui symbolisent respectivement ي et ة [18].

Mot sans voyelles	1ère Interprétation		2ème Interprétation		3ème Interprétation	
كتب	كَتَبَ	il a écrit	كُتِبَ	Il a été écrit	كُتُبْ	des livres
مدرسة	مَدْرَسَةٌ	école	مُدَرَّسَةٌ	enseignante	مُدَرَّسَةٌ	enseignée

Table 1.4 : Ambiguïté causée par l'absence de voyelles pour les mots كتب et مدرسة

1.10. Conclusion

Dans ce chapitre nous avons pu voir le signal de parole depuis sa production jusqu'aux sa perception, on a présenté l'appareil phonatoire humaine ainsi que les différents organes qui entre en jeu dans le processus de production de la parole. On a abordé aussi les paramètres pertinents du signal vocal en se basant sur les paramètres prosodiques, on a étalé la langue arabe et ces propres particularités. Dans le suivant chapitre nous exposerons les bases de reconnaissance automatique de la parole et les différents étages d'un système de reconnaissance automatique de la parole.

CHAPITRE 2
LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

2.1. Introduction

La Reconnaissance Automatique de la Parole a pour but de permettre à un utilisateur de s'adresser oralement à une machine pour des tâches diverses: transcription, commande, traduction, etc. Ce chapitre est une introduction générale à ce domaine pluridisciplinaire. Après en avoir posé les grands principes, nous dressons un bref historique des travaux menés depuis plus d'un demi-siècle. Ce domaine se caractérise actuellement à la fois par le développement d'applications pratiques dans des secteurs variés et par un effort de recherche toujours important, notamment pour augmenter la fiabilité et la robustesse des systèmes dans le cadre de la communication Homme-Machine.

Nous présentons succinctement les différents aspects de ce problème multiforme, nous détaillerons dans ce chapitre la Reconnaissance Automatique de la Parole (RAP) et les fondements théoriques des différents algorithmes qu'elle utilise. L'étude suivra la progression du signal de parole, partant de sa production, passant par son acquisition et finissons sous forme d'une chaîne de mots reconnue.

2.2. Définitions

La reconnaissance automatique de la parole est un exemple typique et difficile du problème général de reconnaissance des formes qui consiste à transcrire un signal continu en une suite discrète symbolique (graphèmes). La Reconnaissance Automatique de la Parole (RAP ou ASR pour "Automatic Speech Recognition" en anglais) est l'opération qui consiste a identifié des

formes d'ondes sonores fournies par un locuteur, et reçu à travers divers senseurs ou capteurs.

La RAP est aussi une technologie informatique permettant à un logiciel d'interpréter une langue naturelle humaine. Elle permet à une machine d'extraire le message oral contenu dans un signal de parole. Cette technologie utilise des méthodes informatiques des domaines du traitement du signal et de l'intelligence artificielle [22]

2.3. Historique de la reconnaissance de la parole

Le tableau 2.1 [23] propose un historique succinct de l'évolution des systèmes de reconnaissance de la parole. Les premiers systèmes de reconnaissance de la parole ne reconnaissent que les mots isolés et nécessitent une phase d'apprentissage longue et fastidieuse. L'apparition dans les années 60 des méthodes numériques et l'utilisation généralisée des ordinateurs apportent un renouveau à ce domaine dont les résultats restent toutefois modestes, quelques 500 mots isolés sont reconnus avec des systèmes dépendants du locuteur . Les difficultés liées en particulier à la parole continue avaient été sous-estimées, telles que la variabilité du signal due au locuteur (état émotionnel ou physique, genre du locuteur, accent), la variabilité due aux conditions acoustiques (type de microphone, bruits...), la variabilité du canal de transmission (téléphone, radio...) et la variabilité due à la langue (discours spontané, hésitations, silences, reprises...). Vers 1970, on s'intéresse à l'apport de contraintes linguistiques dans le processus de décodage automatique de la parole, et les connaissances en microélectronique conduisent à une augmentation de la puissance des ordinateurs.

À partir de ce moment, deux voies de recherche ont été inspectées : la reconnaissance globale et la reconnaissance analytique. Les systèmes à démarche globale ont été conçus à l'origine pour reconnaître directement les mots dans une tâche de reconnaissance limitée. Le contexte est ici dépendant du locuteur en mots isolés dans une ambiance peu bruitée.

Les systèmes à démarche analytique ont été développés pour la plupart pour reconnaître de la parole continue, multilocuteur, à grand vocabulaire et langage peu contraint. Les systèmes Summit sont des exemples de systèmes à démarche analytique. Depuis, les systèmes sont capables de s'adapter à n'importe quel locuteur, ils gèrent la parole continue avec un vocabulaire de plusieurs centaines de milliers de mots, voire illimitée en milieu calme.

Année	Événement
1952	Reconnaissance de chiffres par dispositif électronique câblé Système dépendant du locuteur utilisant les densités de passage par zéro [Davis 1952]
Années 60	Méthodes centi-secondes où une liste d'étiquettes phonétiques est attribuée à chaque 10ms de signal Systèmes dépendants du locuteur
1965	Reconnaissance de phonèmes en parole continue pour le japonais [Doshita 1965]
1968	Reconnaissance de mots isolés (500 mots) [Vicens 1969]
1969	Utilisation d'informations sémantiques et syntaxiques [Vicens 1969, Tubach 1970]
Années 70	Méthodes basées sur la programmation dynamique (DTW : Dynamic Time Warping [Higgins 1991]) Efficace pour des vocabulaire de petites tailles et des systèmes dépendants du locuteur
1971	Lancement du projet ARPA aux USA visant à tester la faisabilité de la compréhension automatique de la parole avec des contraintes raisonnables
1972	Premier appareil commercialisé de reconnaissance de mots isolés (24 mots) : le VIP100 [Hersher 1972]
1976	Fin du projet ARPA : les systèmes opérationnels sont HARPY [Lowerre 1976], HEARSAY II [Lesser 1975] et HWIM [Woods 1976]
Années 80	Méthodes statistiques utilisant les HMMs [Jelinek 1976] Amélioration du taux de reconnaissance SRAP indépendants du locuteur à grands vocabulaires en parole continue
1983	Première utilisation mondiale d'un système à commande vocale à bord d'un avion de chasse
1985	Les systèmes de reconnaissance dépassent le millier de mots reconnus
1986	Lancement du projet japonais ATR utilisant la traduction automatique en temps réel par le téléphone [Fujisaki 1987]
1988	Premières machines de dictée vocale par mots isolés
1989	Premier système de reconnaissance CMU Sphinx [Lee 1989]
Années 90	Méthodes hybrides utilisant les HMMs et les réseaux de neurones [Bourlard 1994, Franco 1992] Systèmes plus robustes au bruit, plus rapides et plus performants
1993	Premier SRAP de parole continue (langue allemande) fonctionnant en quasi temps réel présenté par Phillips à la conférence Eurospeech [Steinbiss 1993]
1993	IBM lance son premier système de reconnaissance vocale sur PC : *Speech Server Series* [Derouault 1993]
1997	IBM lance une machine à dictée vocale en parole continue : *IBM Voice Type-Dictée Personnelle* [Crépy 1997]
Années 2000	Compétition de plus en plus vive des différents laboratoires de recherche et des industriels notamment avec les campagnes d'évaluation NIST[4] Chutes des prix des produits proposés et amélioration notable des performance Démocratisation des produits notamment avec l'arrivée des serveurs vocaux par téléphone Dynamisation de la communication parlée grâce aux nouvelles technologies telles qu'Internet

Table 2.1 : Historique de la reconnaissance de la parole et de ses applications [23]

2.4. Domaines d'application de la RAP

Nous allons cité dans ce qui suit les domaine d'application de la RAP les plus important et dans les quels des SRAP existe déjà ou ils sont en voie de construction, ces domaine sont :

- la sécurité;

- le contrôle d’accès, on peut cité à titre d'exemples:
    - commander une voiture vocalement, sécurisé l'accès à une banques, ou une entreprise vocalement
    - consultation un compte bancaire à distance à travers l'utilisation de téléphone
- le domaine militaire :
    - police criminelle (identification de suspects)
    - filtrage de voix suspectes (avec validation humaine)
    - commandes vocales en navigation aérienne;
- la transcription automatique :
    - adaptation des modèles acoustiques à la voix du locuteur.

2.5. Quelques principaux objectifs de la RAP

La reconnaissance vocale a eux une énorme utilisation surtout dans les services a usage générale tel que le service Télécom (utilisation des téléphones portables) et cela pour atteindre les objectifs suivants :

- améliorer la fiabilité des systèmes tout en passant de monde de fonctionnement indépendant de locuteur vers un autre monde totalement sécurisé et fortement liée au locuteur.

- augmenter l'interactivité des systèmes Hommes-Machines, tout en intégrant le module de la reconnaissance de la voix dans ces systèmes

- rendre la phase de reconnaissance de la parole robuste surtout dans les environnements bruités

- tester l'adaptation de la reconnaissance sur des applications réelles et avec un énorme vocabulaire.

2.6. Exemples des logiciels de reconnaissance

- PhonoLor, est un phonétiseur permettant de transformer la transcription orthographique d'un mot ou d'une suite de mots en une transcription phonétique. Ce logiciel utilise des règles de phonétisation apprises sur un corpus d'exemples.

- Snorri et WinSnoori; Snorri est le logiciel d'étude de la parole développé et amélioré depuis 10 ans. Il est destiné à faciliter le travail du chercheur en reconnaissance de la parole, en phonétique, en perception ou encore en traitement du signal. Les fonctions de base de Snorri permettent de calculer plusieurs types de spectrogrammes et d'éditer le signal de parole de manière très fine (couper, coller, filtrages et atténuations diverses) car le spectrogramme permet de connaître la répercussion acoustique de toutes les modifications. À cela s'ajoute un grand nombre de fonctions destinées à étiqueter phonétiquement ou orthographiquement des signaux de parole, des fonctions destinées à extraire la fréquence fondamentale de la parole, des fonctions destinées à piloter le synthétiseur de Klatt et d'autres à utiliser la synthèse PSOLA. Initialement développé sous Unix et Motif, nous l'avons porté sous Windows et nous le commercialisons depuis 1999 sous le nom de WinSnoori par l'intermédiaire de Babel Technologies (startup située à Mons en Belgique et vendant des logiciels de synthèse et de reconnaissance automatique de la parole) [24].

- Étiquetage de corpus écrits pour la reconnaissance: Il existe des outils d'étiquetage permettant de résoudre syntaxiquement un texte. Ils permet d'affecter à chaque mot d'une phrase sa classe syntaxique en fonction du contexte dans lequel celui-ci apparaît

2.7. Principe de fonctionnement d'un système de RAP

La reconnaissance de parole peut être vue comme une tâche spécifique de la reconnaissance des formes, sa particularité réside dans la forme à reconnaître tel qu'un signal vocal est met à l'entrée de processus de la Reconnaissance de Formes (RF).

La démarche classique d'un système de RF consiste a opérer selon le schéma générale de la Figure 2.1. Ce schéma n'est pas purement linéaire, des interactions peuvent apparaître entre les différents niveaux pour d'éventuelles retours en arrière [25].

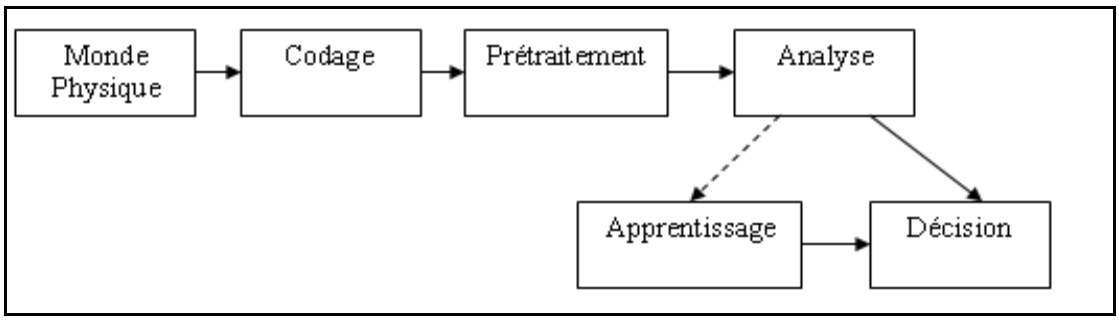


Figure 2.1 : Schéma général d'un système de reconnaissance des formes [25].

- *le monde physique :* on peut l'appeler aussi l'espace des formes, il représente notre monde réel. Les éléments de cet espace sont représentés par un ensemble énorme de propriétés ;

- *le codage :* le traitement informatique de la parole nécessite un signal numérique. Cette nécessité est le but de l'opération de codage, cette dernière consiste à numériser les objets du monde réel, Donc le codage assure le passage du monde physique analogique vers le monde discret numérique ;

- *le prétraitement :* le rôle de cette étape est de supprimer la redondance et réduire le bruit qui est dû aux conditions d'acquisitions à un niveau acceptable et cela grasse à l'utilisation d'un ensemble de filtres ;

- *l'analyse :* l'étape d'analyse représente l'étape la plus importante dans le processus de reconnaissance; Elle sert à extraire un ensemble d'attributs pertinents, discriminants et robustes afin d'être utilisé dans la phase

d'apprentissage pour identifier les objets à reconnaître de manière correcte. En RAP l'analyse vise à extraire à un ensemble d'indices acoustiques (fréquence, intensité, durée) des unités sur lesquelles portera la “décision”, Le choix de ces unités est vaste : phonèmes, diphones, syllabes, mots, etc.

- *l'apprentissage :* on réalité même les machines en besoin d'apprendre; l'apprentissage consiste a crée pour chaque forme en entré une ou plusieurs formes de référence qui sont rangées dans un dictionnaire appelé dictionnaire de références. Le rôle de cette phase d'apprentissage est d'éclairer la décision à l'aide de reconnaissances a priori sur les formes; il existe deux types d'apprentissage supervisé, et non supervisé [25] ;

  - *apprentissage supervisé :* L’apprentissage est dit supervisé si les différentes familles des formes sont connues à priori et si la tâche d’apprentissage est guidée par un superviseur ou professeur, c’est-à-dire le concepteur, celui-ci indique pour chaque forme d'échantillon rentrée, le nom de la famille qui la contienne ;
  - *apprentissage non supervisé :* On l’appelle aussi, suivant l’approche utilisée, classification automatique ou apprentissage sans professeur, Il s’agit, à partir d’échantillons de référence et de regroupement ou de modélisation, de construire automatiquement les classes ou les modèles sans intervention de l’opérateur. Ce type d’apprentissage nécessite un nombre élevé d’échantillons et des règles de construction précises et non contradictoires pour bien assurer la formation des classes.

- *la décision :* La décision ou la classification est une étape de la reconnaissance son rôle est d'identifier la forme ou le signal testé (prévoir l'appartenance de ce signal inconnu a la classe qui lui correspond.) à partir de l'apprentissage réalisé. La prise de la décision ce base sur deux approches [26] :

- *l'approche statistique* (probabiliste) : Elle se base sur l'étude statistique des attributs associe r a l'objet a reconnaître, la prise de décision est de type "associer l'objet a la classe la plus probable" parmi les techniques statistiques existé nous citons :

    - classification automatique ;

    - classification bayésiènne ;

    - classification par discrimination fonctionnelle ;

    - classification par les plus proches voisins connexionniste ;

    - classification stochastique " par chaînes de Markov ".

- *L'approche structurelle* (déterministe) : Elle est sauvant utilisé dans le cas ou l'objet a reconnaître est caractérise par un nombre trot important d'informations, cette approche consiste a trouvé dans cet objet. Les éléments significatifs et leurs relations parmi les méthodes structurelles nous avons :

    - les méthodes syntaxiques ;

    - les méthodes structurelles.

En réalité tout système automatique de reconnaissance vocale se compose généralement de deux principaux modules (Figure 2.2) :

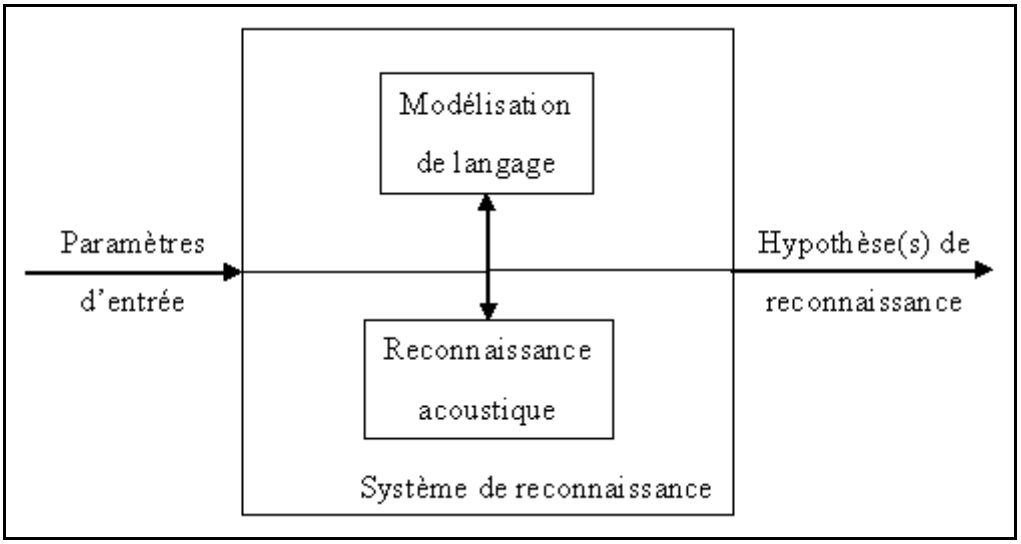


Figure 2. 2 : Description symbolique D'un Système de Reconnaissance de la parole [27].

Le module de modélisation du langage qui permet d'obtenir une information *a priori* sur le positionnement d'un mot dans le signal à reconnaître par différentes techniques de modélisation soit à base de grammaire, soit purement statistique, soit à base d'approches mixtes telles que les grammaires probabilistes [27].

Et le moteur de reconnaissance vocale dont le rôle est de traduire toutes sonores en un texte écrit, en reconnaissance de la parole continue ce module est appelé le module de Décodage Acoustico-Phonétique (DAP), son rôle est de transformer le signal acoustique, en une suite d'unités phonétiques. Les méthodes actuellement les plus performantes dans ce domaine sont fondées sur les HMM.

2.8. Les techniques de la RAP

Parmi les principales méthodes de reconnaissance du signal vocale on a :

2.8.1. Reconnaissance par comparaison à des exemples :

Cette technique est basée sur le principe du calcul de la distance minimale tel que chaque nouvelle forme (chaque nouveau mot à reconnu) est affecté à la classe la plus proche c'est-à-dire la classe qui a un vecteur acoustique qui colle le mieux à celle de mot inconnu (Figure 2.3).

Ce principe de base n'est cependant pas implantable directement : un même mot peut en effet être prononcé d'une infinité de façons différentes, en changeant le rythme de l'élocution. Il en résulte des spectrogrammes plus ou moins distordus dans le temps. La superposition du spectrogramme inconnu aux spectrogramme de base doit dès lors se faire en acceptant une certaine "élasticité " sur les spectrogrammes candidats.

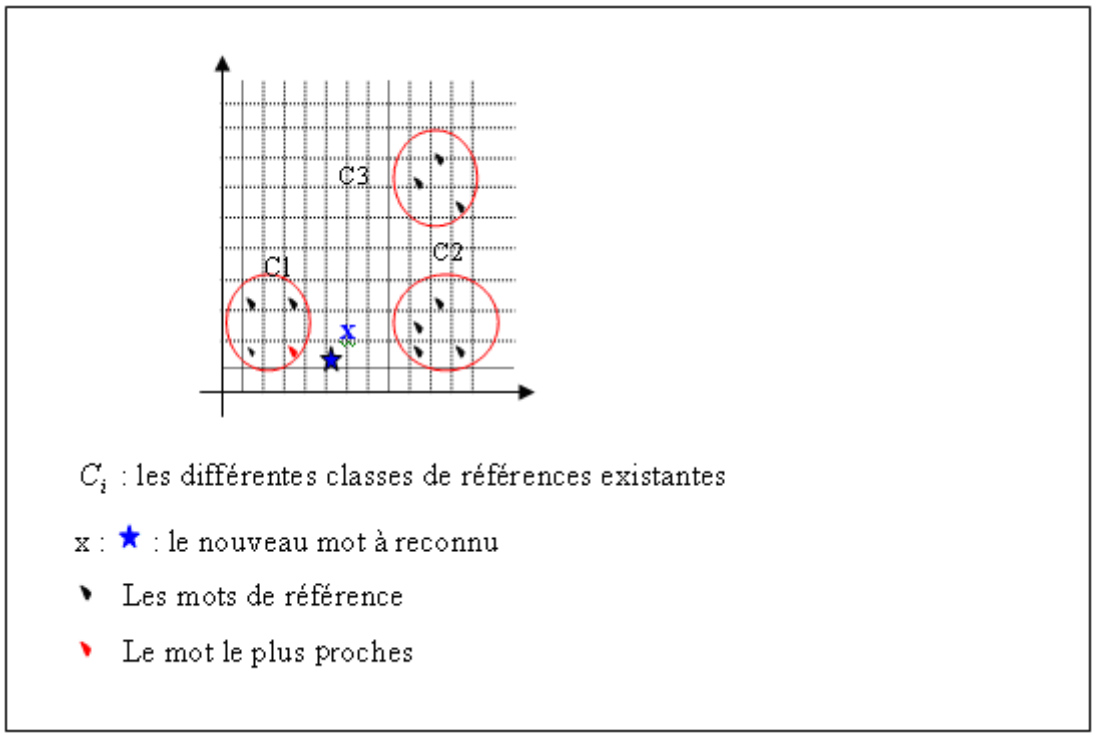


Figure 2. 3 : Principe de fonctionnement de la méthode de reconnaissance par comparaison à des exemples.

Cette notion d'élasticité est formalisée mathématiquement par un algorithme désormais bien connu : l'algorithme DTW [4]. Cet algorithme sera bien détaillé dans le chapitre suivant. Le principe de fonctionnement d'un SRAP basé sur la DTW est présenté dans la Figure 2.4 :

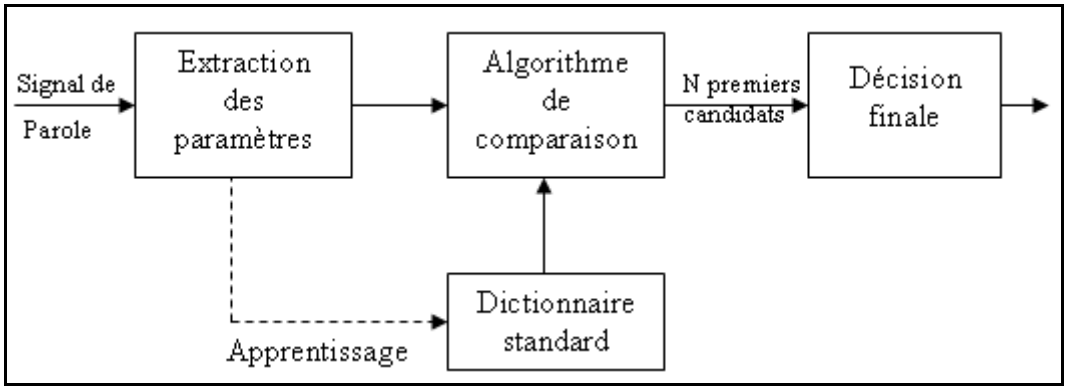


Figure 2. 4 : Principe d'un système de reconnaissance à base de DTW [28].

2.8.2. Reconnaissance par modélisation d'unités de parole

L'inconvenant majeur de la technique de reconnaissance par comparaison à des exemples est le vocabulaire restreint quelle utilise; pour éliminer ce problème est pour utiliser un système de reconnaissance à plus grand vocabulaire et qui s'adaptant facilement à n'importe quelle application, l'idée consiste a donner une représentation abstraite (un modèle) qui est associer à l'opération de reconnaissance (généralement la reconnaissance des phonèmes) (Figure 2.5)

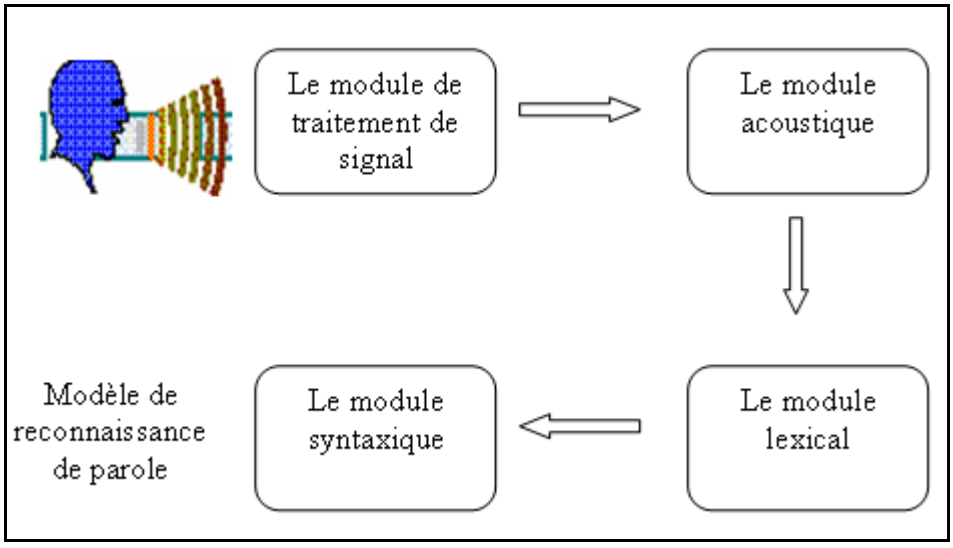


Figure 2. 5 : Reconnaissance par modélisation d'unités acoustiques.

- *le module de traitement de signal :* c'est la phase d'analyse acoustique proprement dite; dans cette étape chaque onde sonore est transformé en un vecteur d'attributs ;

- *le module acoustique :* il est appelé aussi générateur d'hypothèses locales. Pour chaque portion sonore (vecteur acoustique) de 10 ms ce module produire une ou plusieurs hypothèses phonétiques, associées en général à une probabilité. Pour cela il utilise des modèles statistiques (réseaux de neurones artificiels,ou bien des lois statistiques paramétriques) d'unités élémentaires de parole qui sont testé sur plusieurs phrases contenant l'unité sonore associer au vecteur acoustique étudier . cette phase de génération d'hypothèses est toujours accompagner par un module d'alignement temporel (pattern matching, en anglais) qui transforme les hypothèses prises sur chaque vecteur acoustique indépendamment en une décision prise en considérant tout les vecteurs acoustiques grâce à l'utilisation des HMMs. Ce qui nous donne le *modèle acoustique* d'un reconnaisseur quelconque de parole ;

- *le module lexicale :* c'est des simples dictionnaires phonétiques qui contiennent quelques mots de base de la langue étudiée, ces dictionnaires sont souvent utilisés pour forcé le reconnaissance à ne reconnaître que des mots qui existent réellement dans la langue considérée ;

- *le module syntaxique :* ce module consiste a rajouter quelques contraintes syntaxiques au modèle acoustique déjà obtenu.

2.9. Facteurs de complexité

Nous nous proposons ici de recenser les facteurs qui déterminent le degré de complexité d'une tâche de reconnaissance automatique de la parole, avant d'introduire les termes qui servent souvent à qualifier les systèmes qui tentent de résoudre différentes variantes du problème.

En effet, la réalisation d'une machine capable de distinguer quelques mots de commande prononcé isolément par une personne donnée et celle

d'un système (encore fort hypothétique!) à qui n'importe qui pourrait s'adresser comme à un interlocuteur humain, pour lui dicter un texte quelconque, ou lui demander une information, il y a manifestement un monde!

Les facteurs déterminant sont les suivants :

- *Reconnaissance mono-locuteur, ou non*

  La reconnaissance de la parole pour la voix d''un seul locuteur n'est déjà pas un problème élémentaire, en raison de la variabilité intra-locuteur, inhérente au processus humain de production. Cette variabilité est en réalité liée aux variation du débits d'élocution, à l'émotion, au stress, aux rhumes, etc. la reconnaissance pour plusieurs locuteurs voir indépendante du locuteur est un problème plus difficile, dans la mesure où à la variabilité "intra" s'ajoute la variabilité inter-locuteur, liée rappelons le, aux caractéristiques anatomiques, aux habitudes linguistiques, accents régionaux, etc. ;

- *Reconnaissance de mots isolés ou de parole continue*

  Si le locuteur marque une pause après chaque mots de l'énoncé, la complexité du problème est réduite puisque les frontière des mots sont alors disponibles (contrairement au cas de la parole continue), ce qui limite beaucoup la combinatoire pour accéder à un lexique, par exemple. De plus les mots parlés peuvent alors éventuellement être considérés comme des entités globales et non comme une suite d'entités élémentaire ;

- *Taille du vocabulaire*

  Toute chose égales d'ailleurs, il sera plus difficile de travailler sur un vocabulaire très étendu (quelques milliers ou dizaine de milliers de mots) que sur un vocabulaire très restreint (quelque dizaine de mots). La taille du vocabulaire est cependant un paramètre insuffisant, un ensemble de mots très différents les uns des autres étant plus facile à traiter que des mots proches phonétiquement. La prise en compte de la syntaxe du

langage produit par l'utilisateur (et éventuellement de sa sémantique, de sa pragmatique dans le cadre d'une application) sera plus facile pour un langage rigide, très contraint, que si toute la souplesse de la langue naturelle parlée est rencontrée ;

- *Environnement*

Indépendamment de ce qui précède, l'environnement acoustique et les conditions de prise de son constitue un facteur important : la présence de bruit, même stationnaire, dégrade fortement les performances des systèmes de reconnaissance. De plus, si ce bruit est intense, il induit une augmentation de variabilité chez le locuteur (effet lombard).

2.10. Type de méthodes et système

Nous proposons maintenant plusieurs critères de classification des méthodes et systèmes de reconnaissance. Ces critères ne sont pas totalement indépendants, mais aucun ne se réduit exactement à un autre, et ils traduisent plutôt des points de vues différents.

- *Reconnaissance et compréhension*

Il s'agit là de deux objectifs différents que l'on peut assigner à un système : la reconnaissance (au sens strict) conduisant à une application du type dictée automatique (la "machine à écrire phonétique"), tandis que la compréhension automatique cherche à accéder à la signification de l'énoncé parlé, par exemple dans le cadre d'un dialogue homme-machine. Cette distinction ne se manifeste vraiment que dans le cas de système assez ambitieux et généraux (pour un langage de commande simple par mots isolés, ou en entrée de données, identification vaut "compréhension") ;

- *Système "auto-organisateurs" et "fondés sur des connaissance"*

Un système peut utiliser des méthodes mathématiques pour extraire de données représentatives (échantillon d'apprentissage) les structures et

paramètres qu'il utilisera pour effectuer ses connaissances. Il est alors dit "auto-organisateur". Cela ne signifie pas qu'il soit totalement ignorant des connaissance humaines (qui aurons bien probablement aidé le concepteur à définir la meilleur façon de mettre en œuvre les dites méthodes mathématiques sur un type spécifique de données), mais simplement qu'il ne les utilise pas explicitement. La programmation dynamique et la modélisation Markovienne relèvent typiquement de cette rubrique ;

Une autre approche est de tenter d'utiliser explicitement dans le système les connaissance que possèdent consciemment les experts humains qui ont étudié la parole, et ce à plusieurs niveaux : acoustique, phonétique et linguistique. On parle alors de système "fondés sur des connaissances". De tels systèmes sont assez naturellement amenés à mettre en œuvre des traitements de types intelligence artificielle (coopérant avec les traitements classiques de traitement de signal et de reconnaissance des formes) ;

- *Reconnaissance globale par mots et reconnaissance analytique*

Certaines méthodes consistent à comparer entre elle des entités de la taille des mots que l'on est censé identifier, sans les décomposer en entités plus élémentaires. On parle alors de reconnaissance globale. D'autres au contraire utilise à un stade intermédiaire des entités élémentaires (phones, diphones, demi-syllabes…), qui permettent de décrire (via un lexique le plus souvent) les entités de niveau supérieur : c'est la reconnaissance analytique.

Cette distinction mériteraient plus de subtilité; par exemple, les unités élémentaires de la reconnaissance analytique n'ont pas nécessairement une définition phonétique, mais peuvent être des unités "ad hoc" (définie par exemple par quantification vectorielle). De toutes façons la distinction analytique/globale a perdu une partie de sa netteté depuis que les méthodes globales, limité au début à la reconnaissance de mots isolés, ont été généralisés aux mots connectés, et par ce biais peuvent visé la

parole continue (syllabes connectées…), tandis que le traitement de très grand vocabulaire, même en élocution isolé, paraît requérir des méthodes analytique.

Typiquement, les systèmes actuellement commercialisés pour la reconnaissance de mots sont globaux et auto-organisateurs, tandis que sont à l'étude, pour des interaction homme-machine complexe, des systèmes de compréhensions, analytiques, au moins partiellement fondé sur des connaissances. Mais se ne sont pas les seules combinaisons possibles.

2.11. Applications de la RAP

Un SRAP peut être vu comme un module appartenant à un système plus important. Les hypothèses fournies par le SRAP sont généralement utilisées dans diverses applications comme la commande vocale, les systèmes de dialogue (demande d'informations), la dictée vocale, la transcription automatique, la traduction, ou encore pour l'indexation de données audio et audiovisuelles. Certains systèmes comme les systèmes de demande de mot de passe utilisent en plus un module de reconnaissance du locuteur.

2.11.1. Commande vocale

Il s'agit ici de systèmes le plus souvent dépendants du locuteur pour la reconnaissance de mots isolés. De nombreux systèmes à commande vocale sont utilisés dans des avions de chasse, des automobiles, pour manoeuvrer des objets à distance ou encore pour l'aide aux personnes handicapées. En effet, dans des endroits exigus comme la cabine de pilotage d'un avion, la parole permet au pilote ou au conducteur de disposer un nouveau moyen d'interaction avec sa machine sans pour autant gêner son attention visuelle. Pour améliorer la productivité humaine, les systèmes à commande vocale permettent par exemple d'effectuer l'inventaire du stock d'une entreprise commerciale de manière plus efficace que de manière écrite.

2.11.2. Systèmes de dialogue

Ce sont des systèmes multilocuteurs qui, en plus d'un module de reconnaissance de la parole, incluent des modules de compréhension, de synthèse de la parole, de génération de phrases et d'interrogation de bases de données. La figure 2.6 montre le fonctionnement d'un tel système. La plupart de ces systèmes fonctionnent à partir du téléphone et permettent d'orienter l'utilisateur à travers une base de données comme les réservations de billets de train.

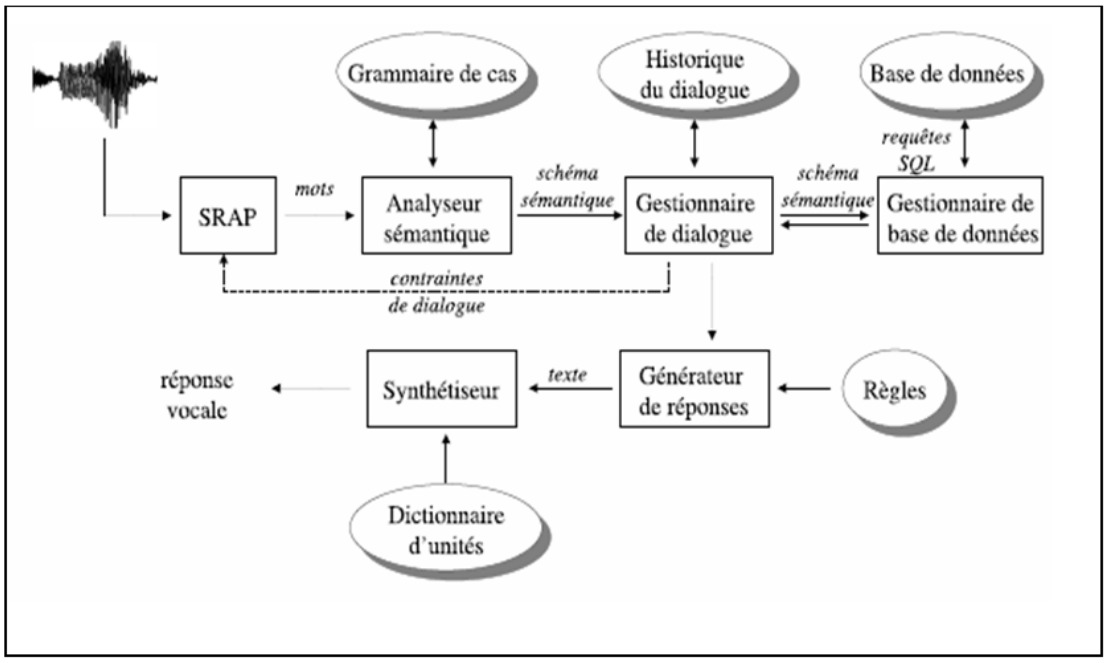


Figure 2.6 : Système de dialogue [23].

2.11.3. Dictée vocale

Pour le grand public et les professionnels, on trouve bon nombre de logiciels facilitant la prise de notes grâce à la transcription de la parole de l'utilisateur. Pour cette application, le locuteur est connu (système dépendant du locuteur) et les conditions de prise de son sont généralement optimales du fait de la proximité du micro dans un environnement assez peu bruité. Elle nécessite également un long temps d'apprentissage afin que le système s'adapte à la voix, notamment aux accents régionaux, aux défauts d'élocution et de prononciation... Pour la parole spontanée, les performances de tels

systèmes se situent aux alentours de 14%. Pour des textes lus, le taux d'erreur est d'environ 7% sur de l'anglais américain, des résultats similaires ont été observés pour le français (campagne AUPELF) et l'allemand [23].

2.11.4. Traduction automatique

Ces systèmes tentent de pallier la barrière de la langue. Des applications comme la traduction de mails oraux, de cours magistraux en direct ou tout simplement la traduction instantanée d'un locuteur étranger est envisagée par ce domaine. La figure 2.7 montre le diagramme d'un traducteur automatique parole- arole.

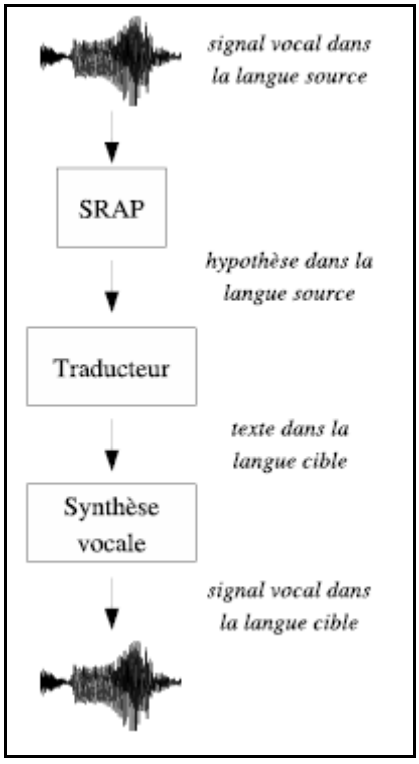


Figure 2.7 : Diagramme d'un traducteur parole-parole [23]

Le signal vocal en langue source est transcrit au moyen d'un SRAP. Ce texte est traduit dans la langue cible grâce au composant de traduction. Ensuite, la synthèse vocale permet de transformer ce texte en signal vocal dans la langue cible. La traduction automatique de la parole nécessite de résoudre plusieurs problèmes : d'une part reconnaître la parole continue prononcée par un locuteur quelconque, puis en comprendre le sens pour

générer un énoncé dans la langue cible et enfin le synthétiser avec une formulation et une voix les plus naturelles possibles. Dans le cadre de la traduction parole-parole, le consortium C-STAR (Consortium for Speech Translation Advanced Research) fait coopérer plusieurs laboratoires (Japonais, Coréen, Anglais, Français, Allemand et Italien) et permet le regroupement de systèmes de traduction d'une langue à un autre [23]. Ce consortium permet de dynamiser les recherches de chacun des laboratoires et permet également les interactions entre chacun des partenaires. Pour le moment, ce consortium est centré autour des domaines touristiques tels que la demande d'informations, de réservations et de planning.

2.11.5. La reconnaissance du locuteur

La reconnaissance automatique du locuteur (RAL) vise à déterminer si un échantillon de voix a été prononcé par une personne donnée. Elle peut être scindée en deux catégories :

- *Identification du locuteur :* parmi un ensemble de locuteurs connus, il s'agit de rechercher l'identité du locuteur possédant la référence la plus proche du signal vocal donné ;

- *Vérification du locuteur :* étant donné un signal vocal et une identité proposée par un locuteur, il s'agit d'accepter ou de rejeter l'hypothèse que le locuteur considéré l'ait prononcé.

Le diagramme 2.8 représente une application d'authentification du locuteur comme dans les applications de serrure vocale. Le système connaît les locuteurs autorisés à ouvrir la serrure alors que des imposteurs peuvent tenter de la franchir. Le système d'authentification du locuteur peut par exemple être obtenu comme sur cette figure en mettant en cascade un système d'identification du locuteur avec un système de vérification.

Les applications liées à la reconnaissance du locuteur sont de l'ordre du contrôle d'accès, de la vérification de présence, de la protection d'équipements, de l'authentification, de la personnalisation d'informations...

Certaines applications comme le contrôle d'accès font appel à des mots de passe où la vérification du locuteur est dépendante de la reconnaissance d'un mot isolé. Ici, l'environnement de prise de son est généralement isolé du bruit. L'énoncé de chaque utilisateur a été appris et est stocké dans une base de données de références vocales.

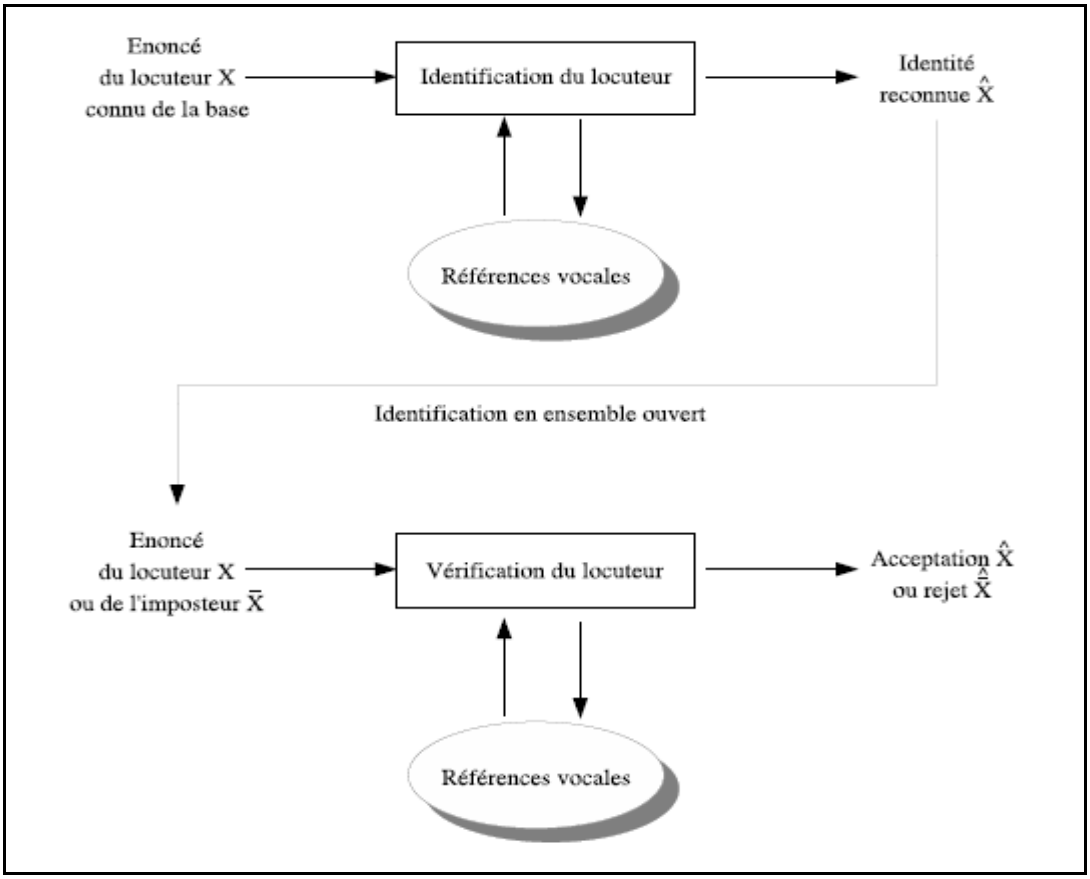


Figure 2.8 : diagramme représentant un système d'identification du locuteur (en haut) couplé avec un système de vérification du locuteur (en bas) [23].

Des méthodes de comparaison sont enclenchées lors de la vérification du mot de passe telles que les méthodes DTW, des méthodes basées sur les HMMs, ou encore des méthodes de modélisation statistique combinée à des informations temporelles (GDW : Gaussian Dynamic Warping) [23].

2.12. Conclusion

Dans ce chapitre nous avons discuté des principaux techniques et méthodes de la reconnaissance automatique de parole (RAP), nous avons vu aussi que la RAP est une tâche spécifique de la reconnaissance des formes et on a détaillé sont processus de fonctionnement. Différentes applications du traitement de la parole ont été décrites, permettant de survoler les différents domaines qui sont susceptibles d'utiliser les sorties d'un SRAP. Dans le chapitre suivant on abordera les techniques d'analyse vocale qui représentent le maillon de base des systèmes de reconnaissance vocales et qui ont pour rôle d'extraire le vecteur représentant de chaque onde sonore.

CHAPITRE 3
LES TECHNIQUES D'ANALYSE VOCALE

3.1. Introduction

Le premier étage d'un système de reconnaissance de la parole analyse et paramétrise le signal vocal, en vue de minimiser la quantité d'informations nécessaires à la séparation des éléments du vocabulaire. Les méthodes d'analyse ont pour but d'extraire du signal vocal les informations jugées pertinentes afin de les utiliser dans l'étage de reconnaissance ou de décision d'un système de reconnaissance.

Par conséquent, on peut dire que les différents traitements effectués sur le signal de parole ont pour but :

- d'en donner une représentation plus concise du signal, en conservant les paramètres les plus pertinents vis-à-vis de la tâche à accomplir.

- d'extraire des indices permettant de segmenter la parole et de classifier les segments acoustiques obtenus.

Nous verrons dans cette partie les techniques d'analyse les plus répandues : spectrogramme, analyse par prédiction linéaire (LPC) et l'analyse Cepstrale (MFCC). Ensuite nous tournons le voile vers les méthodes de reconnaissance en décortiquant deux méthodes les plus utilisés dans ce domaine, à savoir : la comparaison dynamique (DTW) et les modèles de Markov cachés (HMM).

3.2. Définitions

L'Analyse acoustique est une partie importante dans le traitement que subit le signal sonore pour pouvoir réaliser un système de haute qualité de synthèse, de compréhension, ou de reconnaissance de la parole [29].

Le signal de parole démontre une très grande variabilité. Il est en effet peu probable de mesurer deux signaux de parole totalement semblables même si des mots identiques sont prononcés par le même locuteur. Cette variabilité rend le processus de traitement automatique de parole excessivement complexe. Le rôle du module d'analyse acoustique est de traiter le signal de parole de manière à réduire cette variabilité. Durant cette opération l'ordinateur essaie de construire une représentation formelle du message d'entrée, oral ou écrit, plus facile à manipuler que le message original [30].

3.3. Les différents types d'analyse vocale

L'analyse de la parole peut être envisagée de deux manières. La première consiste à étudier l'évolution du contenu spectral du signal. Ce type d'approche nécessite d'extraire les caractéristiques fréquentielles du signal à intervalle régulier. Nous nommerons cette approche "description acoustique". Le second type d'analyse, nommée "description segmentale", consiste à découper le signal en segments de taille variable. Ces segments peuvent avoir un sens phonétique. On cherchera par exemple à délimiter les unités linguistiques comme les phones, les diphones, etc. Cependant ce type de segmentation nécessite la connaissance a priori du texte à l'origine du signal. Des méthodes de segmentation automatique ont été développées afin de résoudre ce problème ; c'est le cas du le système ALISP 1 [31]. L'Analyse vocale consiste à tirer à partir du signal vocal un ensemble de paramètres pertinents, discriminants et robustes susceptibles de le représenter.

3.4. Les techniques d'analyses

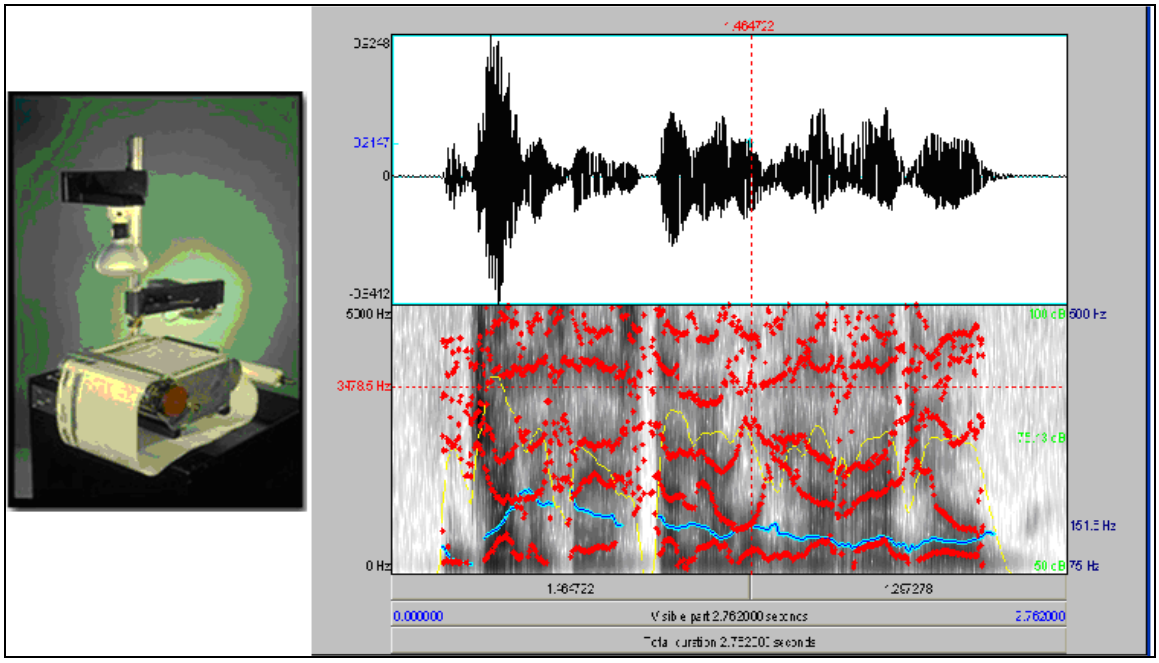
Nous présenterons les techniques d'analyse rencontrées dans la littérature en vue de classifier les sons respiratoires. Ces différentes techniques peuvent être subdivisées en trois groupes :

- analyse par spectrogramme ;
- analyse cepstrale ;
- prédiction linéaire.

3.4.1. L'analyse par spectrogrammes

Le spectrogramme est une représentation 3D (le temps, la fréquence, et l'amplitude) qui associer à chaque instant t d'un signal le spectre de fréquence qui lui correspond (Figure 3.1). Le temps et la fréquence sont représentés sur l'axe

horizontal et vertical, tandis que l'amplitude d'une fréquence particulière à un temps donné est suggérée par le degré de noircissement de l'affichage. Il présente sous une forme simple l'essentiel de l'information portée par le signal vocal car il permet à l'œil expérimenté de certains phonéticiens de retrouver le contenu du message parlé qui est une propriété non évidente pour l'audiogramme qui renseigne peu sur le timbre des séquences sonores.



(a) (b)

Figure 3.1 : (a) L'appareil qui génère le spectrogramme[32], (b) Spectrogramme de la phrase /جلس يستمع إلى ألراديو/.

Donc on peut réduire la description d'un son par trois grandeurs physiques affichés par le spectrogramme : la fréquence (Hz), la durée (s) et l'amplitude (ou l'énergie) (dB). Par exemple, un son de parole simple, c'est-à-dire sinusoïdal telle que la voyelle [fatha] de l'Arabe Standard qui se situe en milieu de mot est complètement décrite par les valeurs [29]:

F_0 = 144.4488 Hz, t = 0.0865 s, A = 83.7061 dB.

Et un autre son complexe (Bruité) sera défini, par exemple, de la manière suivante :

- t = 5 s ;
- F_0 = 100 Hz et A_0 = 70 dB (pitch) ;
- F_1 = 200 Hz et A_1 = 65 dB (2^e harmonique) ;
- F_2 = 300 Hz et A_2 = 50 dB (3^e harmonique).

Cela signifie que les trois valeurs temps, fréquence, et énergie sont les paramètres pertinents donc une meilleure analyse consiste à les représenter de manière claire et avec exactitude. L'une des représentations possibles est d'associer deux à deux ces trois grandeurs et de tracer les graphes de ces associations, on obtient les trois plans suivants :

- dynamique (temps, amplitude) ;
- du spectre (fréquence, amplitude) ;
- mélodique (temps, fréquence).

Le spectrogramme est l'une des méthodes d'analyse qui assure une représentation tridimensionnelle de signal de parole tel que (Figure 2.1) :

- l'axe vertical représente la fréquence du son en Hz ;

- l'axe horizontal représente l'évolution temporaire du son ;

- le degré de noircissement représente l'intensité (l'énergie) en dB du son.

L'objectif principal de spectrogramme est de connaître l'évolution temporelle du spectre de parole. Pour assurer cet objectif, Il faut décomposer l'onde acoustique du son en ondes sinusoïdales de différentes fréquences au moyen d'une transformée de Fourier.

Il existe deux représentations possibles pour un spectrogramme, la première en *bande étroite* et la deuxième en *bande large*. La différence essentielle entre les deux réside dans le choix des paramètres qui nous intéressent [29] :

- un spectrogramme à Bande Large (BL) offre une meilleure résolution fréquentielle et permet de visualiser clairement l'évolution formantique des sons, mais il correspond à une mauvaise analyse temporelle (Figure 3.2. (a)). La classe des occlusives représente le meilleur exemple adapté à cette présentation ;

- inversement, un spectrogramme à Bande Etroite (BE) offre une bonne résolution au niveau temporel, mais l'analyse fréquentielle est moins fine (Figure 3.2. (b)). Ce type de spectrogramme est souvent utilisé dans l'étude de l'intonation ainsi que dans l'analyse des fricatives.

Ces deux représentations sont complémentaires. En effet, le spectre à bande large possède une meilleure définition temporelle et permet d'identifier les fréquences de résonance du conduit vocal (larges bandes foncées) plus précisément que sur le spectre à bande étroite. Cependant celui-ci, grâce à sa meilleure définition fréquentielle, fait apparaître les harmoniques (fines bandes foncées). Il est donc plus utile pour étudier l'influence de la source glottale.

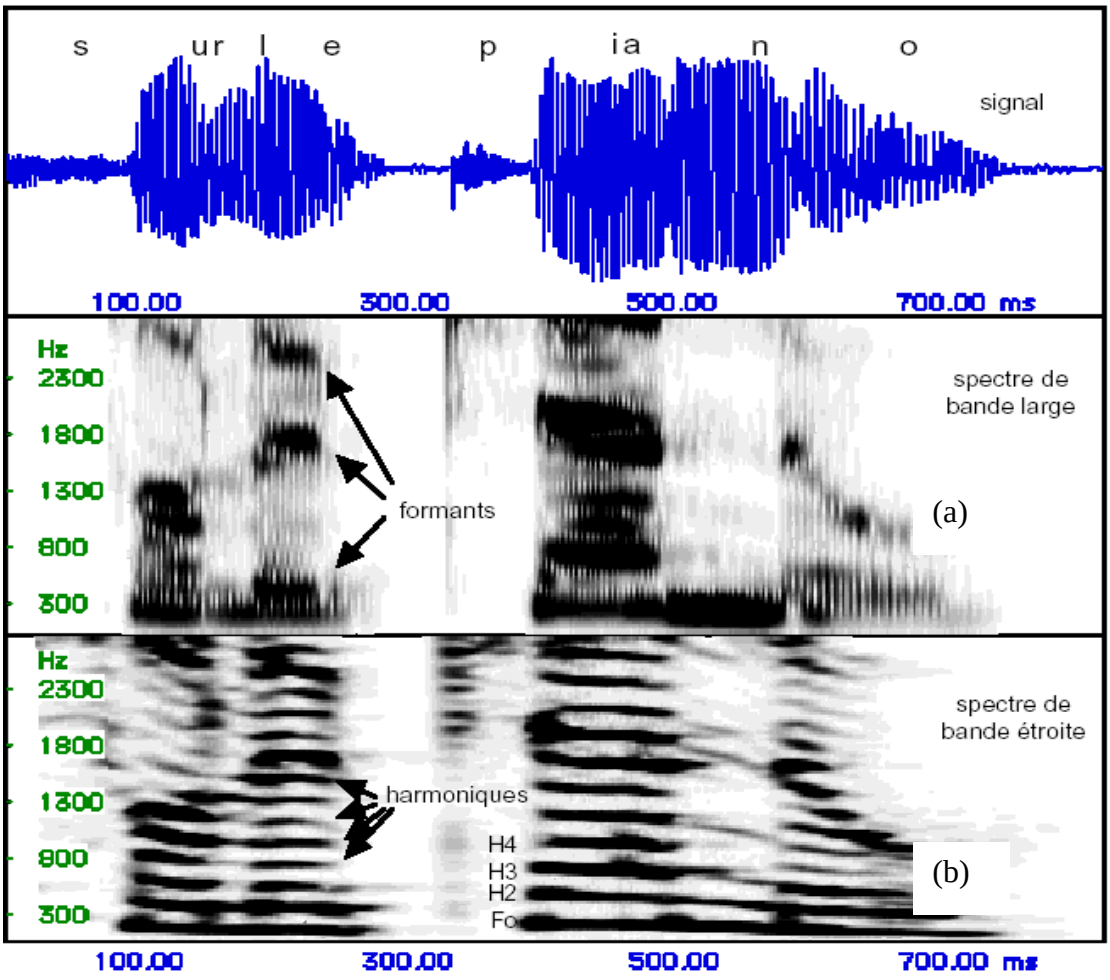


Figure 3.2 : Différentes analyses en BL et BE de l'onde sonore correspondant à l'énoncé " Sur le piano" [29].

3.4.2. L'analyse Cepstrale (MFCC)

L'analyse acoustique traite le signal et en extrait les vecteurs acoustiques qui seront utilisés pour des phrases suivantes. Dans cette étape le signale de parole est transformé en une séquence de vecteurs acoustiques pour diminuer la redondance et la quantité de données à traiter. Pour améliorer la qualité du signal vocal, on peut appliquer en plus des algorithmes pour réduire le bruit. Puis une analyse spectrale

par la transformation de Fourier discrète, est effectuée sur une trame de signal (généralement de taille 10 ou 20 ms) [28].

Dans cette trame, le signal vocal est considéré comme suffisamment stable et on en extrait un vecteur de paramètres, que sont dans notre travail des MFCC (Mel-scale Frequency Cepstral Coefficients). L'utilisation très fréquente des MFCC peut s'expliquer de façon pragmatique par le fait qu'ils constituent la paramétrisation *(état-de-l'art)* de la majorité des systèmes de traitement automatique de la parole acoustique et qu'ils ont démontré leur efficacité que ce soit en transcription de la parole ou en vérification du locuteur [40].

Dans cette partie, nous montrons comment obtenir des vecteurs acoustiques (les paramètres pour les phases d'apprentissage et de test de reconnaissance). Les paramètres les plus utilisés sont les MFCC que nous utiliserons dans nos expériences. La figure 3.3 montre le processus d'extraction des MFCC :

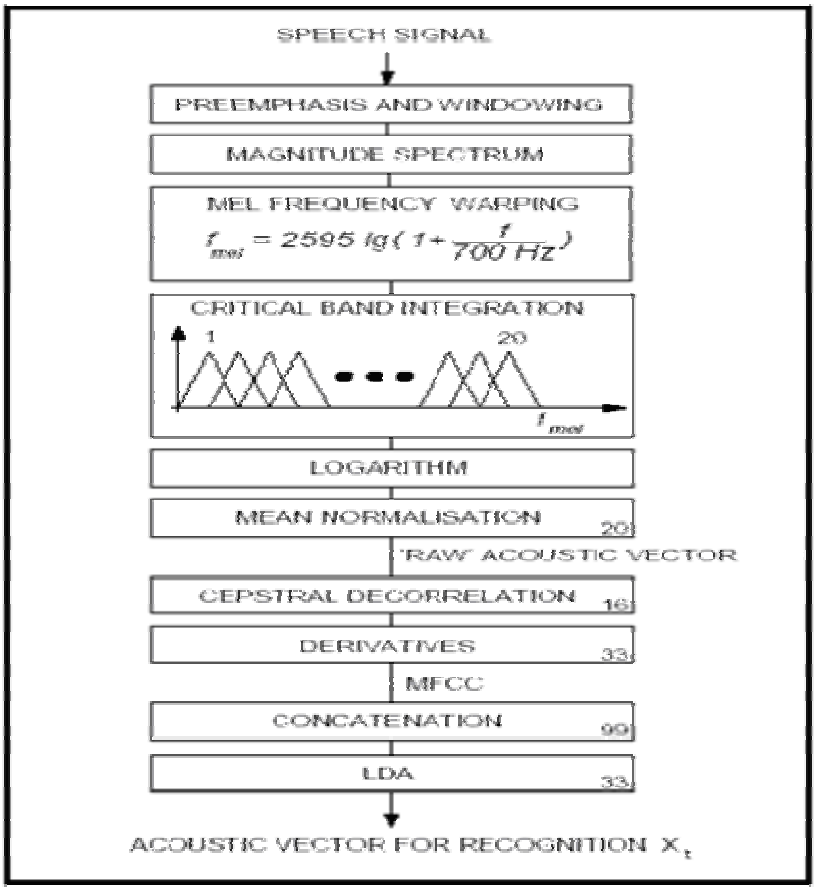


Figure 3.3 : processus d'extraction des coefficients MFCC [28]

La motivation d'extraction de coefficients MFCC est leur correspondance avec la réponse en fréquence d'une oreille humaine. Pour transformer une fréquence linéaire en une fréquence Mel, on peut utiliser la formule de transformation suivante:

$$\mathrm{B}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (3.1)$$

Où *f* est la fréquence en Hz, *B(f)* est la fréquence mel-échelle de *f*. Il existe plusieurs approximations pour la fonction de correspondance entre la fréquence Mel et la fréquence réelle mais la formule ci-dessus est la plus utilisée. Le processus de calcul des coefficients Mel est détaillé par le schéma de la figure suivante :

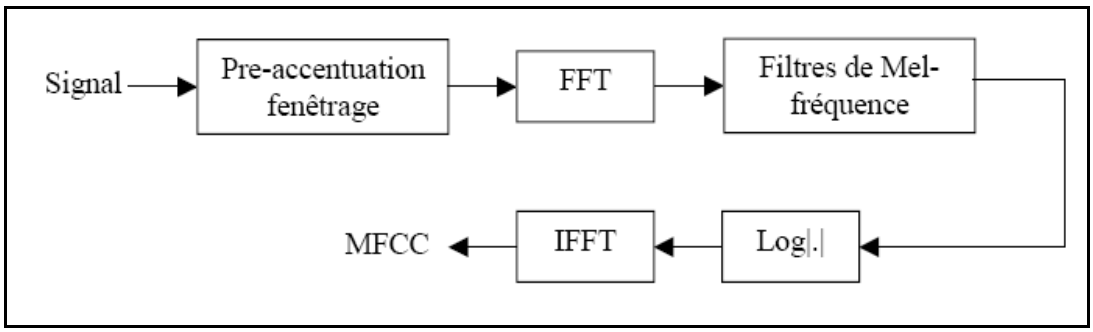


Figure 3.4 : calculs des MFCC [28]

Les signaux *s[n] 0 ≤ n ≤ N-1* sont transformés dans le domaine fréquentiel par la transformée de Fourier discrète FFT. Puis le spectre du signal est multiplié par une suite des filtres triangulaires dont les bandes passantes sont de même taille dans l'échelle Mel.

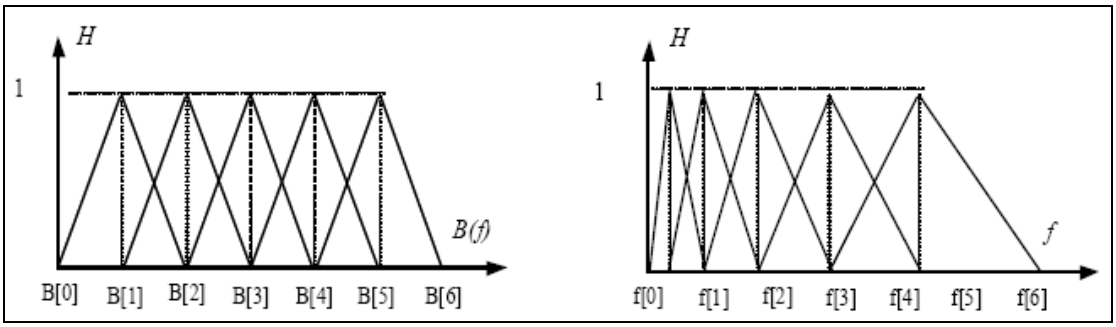


Figure 3.5 : les filtres triangulaires passe-bande en Mel-Fréq (B(f)) et en fréquence (f) [28]

On peut calculer les points frontières *B[m]* des filtres en mel-fréquence ainsi :

$$\mathrm{B}[m] = \mathrm{B}(f_1) + m\frac{\mathrm{B}(f_h) - \mathrm{B}(f_l)}{M+1}, 0 \le m \le M+1 \qquad (3.2)$$

On doit calculer les points *f[m]* correspondants dans le domaine de fréquence réelle :

$$f[m]=\left(\frac{N}{M}\right)\mathrm{B}^{-1}\left(\mathrm{B}(f_1)+m\frac{\mathrm{B}(f_h)-\mathrm{B}(f_l)}{M+1}\right) \qquad (3.3)$$

Puis on détermine tous les coefficients de chaque filtre :

$$H_m[k]=\begin{cases}0\rightarrow k\leq f[m-1]\\ \dfrac{k-f[m-1]}{f[m]-f[m-1]}\rightarrow f[m-1]\leq k\leq f[m]\\ \dfrac{f[m+1]-k}{f[m+1]-f[m]}\rightarrow f[m]\leq k\leq f[m+1]\\ 0\rightarrow k\geq f[m+1]\end{cases} \qquad (3.4)$$

Ensuite, on multiplie toutes les énergies de *S[k]* avec les coefficients $H_m[k]$ et on calcule leur logarithme :

$$E[m]=\log\left[\sum_{k=0}^{N-1}|S[K]|^2 H_m[K]\right] \qquad 0 \leq m < M \qquad (3.5)$$

On a maintenant un jeu de *M* paramètres *E[m] 0 ≤ n ≤ m-1*, où *M* est le nombre de filtres triangulaires. Enfin, les coefficients Cepstraux de mel-fréquence (MFCC) peuvent être obtenus par une transformée de Fourier inverse à partie des coefficients de sortie des filtres. La formule de calcul est la suivante :

$$c[n]=\sum_{m=0}^{M-1}E[m]\cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{M}\right) \quad \text{avec } 0 \leq n < M \qquad (3.6)$$

Le schéma général d'extraction des coefficients MFCC peut être résumé dans la figure suivante :

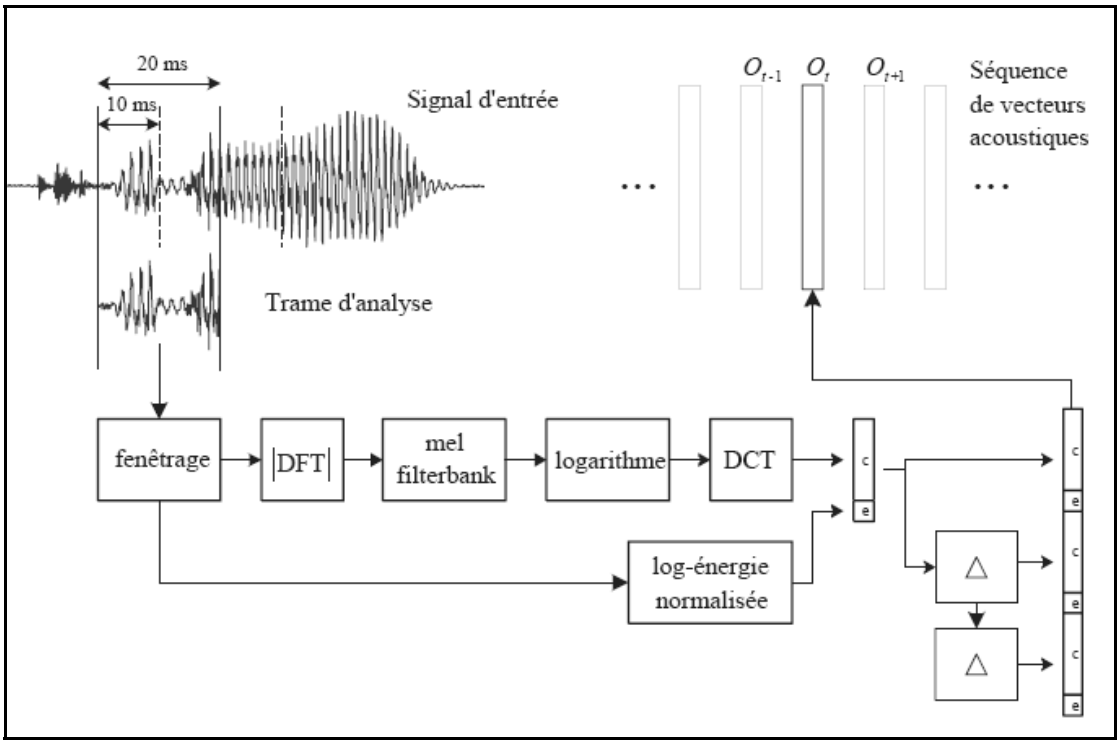


Figure 3.6 : Analyse acoustique par la représentation MFCC [32]

La calcul des MFCC s'effectue ainsi [41] :

- on choisit le nombre *M* de coefficients MFCC, typiquement de l'ordre de 10 à 20 pour la voix ;
- on définit *N* > *M* filtres triangulaires d'importance énergétique équivalente (et donc de largeur de bande variable) sur l'´echelle mel ;
- on pondère la portion de signal par une fenêtre d'analyse (Hamming) pour minimiser la distorsion spectrale avant d'en calculer la TFD ;
- on calcule les énergies du signal analysé X_k (k = 1..N) en sortie des filtres mel;
- on considère la séquence X_k comme un signal temporel discret dont les ondulations (les formants) sont décrites par les coefficients cosinusoidaux de la série de Fourier de ce signal :

$$MFCC_i = \sum_{k=1}^{N} X_k * \cos\left[i\left(k - \frac{1}{2} \right) \frac{\pi}{N} \right]; i = 1..M \qquad (3.7)$$

- on récupère les *M* premiers coefficients $MFCC_i$. En effet, les premiers coefficients prennent en compte les reliefs basse fréquence du spectre

(l'enveloppe spectrale), quand les coefficients d'ordre supérieur décrivent des variations plus rapides dépendantes du pitch et des raies spectrales.

La méthode d'extraction des MFCC est une des méthodes de calcul des vecteurs acoustiques les plus populaires dans le domaine de reconnaissance automatique de la parole. Cette méthode a la capacité de capturer des caractéristiques phonétiques importantes de la parole. Donc, nous avons également décidé de l'utiliser dans notre contexte d'applications et on a choisi un ensemble de 12 coefficients.

3.4.3. Le codage prédictif linéaire

Le codage LPC (*Linear Predictive Coding*) utilise le même principe de production de la parole que ceux utilisés par les vocodeurs traditionnels. Il consiste à synthétiser des échantillons à partir d'un modèle d'un système de production vocal et d'une excitation. Cette dernière est représentée par un signal qui est soit une sinusoïde pour les sons voisés, soit un bruit blanc pour les sons non voisés (Figure 3.7, le gain permet d'ajuster l'énergie du signal synthétisé à celle du signal original)

.

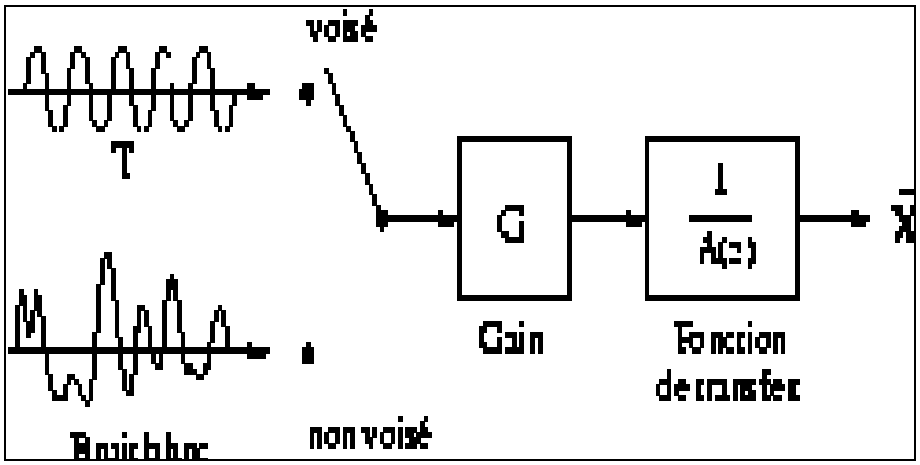


Figure 3.7 : principe du codage LPC

Cette modélisation est appelée prédictive linéaire puisqu'elle correspond à une régression linéaire entre le signal d'excitation et le signal vocal produit, et elle est représentée par la formule suivante :

$$S(n) + \sum_{i=1}^{p} a(i)s(n-i) = G(n) \qquad (3.8)$$

Ce type de codage prédictif linéaire est utilisé pour les téléphones mobiles numériques GSM pour réduire le débit de transmission de parole entre les différentes stations qui constituent ce réseau de téléphonie. En pratique dans la transmission d'informations entre deux stations téléphoniques le signal de parole est échantillonné au rythme de 8000 échantillons par seconde qui correspond presque à un débit de 64 kbits/sec. Il faut ramener ce débit à 13 Kbits/s, dans ce cas la technique retenue est celle de la prédiction linéaire.

Cette technique est fort utilisée dans le codage de parole. Pour la reconnaissance de la parole, elle se justifie en considérant le modèle de parole simplifié source/conduit. Dans ce modèle, le signal de parole γ vaut:

$$\gamma_n = \sum a_k \cdot \gamma_{n-k} + Ge_n \qquad (3.9)$$

$$\mathrm{Y}(z)\left(1 - \sum a_k \cdot z^{-k}\right) = GE(z) \qquad (3.10)$$

La fonction de transfert du conduit est donc égale à :

$$\mathrm{H}(z) = \frac{\mathrm{Y}(z)}{\mathrm{E}(z)} = \frac{G}{1 - \sum_k a_k \cdot z^{-k}} \qquad (3.11)$$

L'analyse par prédiction linéaire conduit à estimer les coefficients de la fonction de transfert du conduit vocal. Elle estime l'échantillon courant $\check{y}_n$ en fonction des *k* échantillons précédents :

$$\check{y}_n = \sum_{k=1}^{K} a_k \cdot y_{n-k} \qquad (3.12)$$

Soit :

$$e_n = y_n - \check{y} \qquad (3.13)$$

L'erreur commise en estimant y_n par $\check{y}$. Le but est de trouver les coefficients a_k minimisant :

$$\mathrm{E} = \sum_n e_n^2 \qquad (3.14)$$

En dérivant l'erreur de prédiction E par rapport aux a_k on obtient :

$$\frac{\partial \mathrm{E}}{\partial a_1} = -2\sum_n y_n - y_{n+1} + 2\sum_n \left(\sum_{k=1}^{K} a_k - y_{n-k}\right) \cdot y_{n-1} \qquad (3.15)$$

Après réorganisation des termes, et en posant :

$$\chi_{\kappa,1} = \sum_n y_n - \kappa \cdot y_{n-1} \qquad (3.16)$$

On obtient les équations de Yule-Walker :

$$\chi_{1,0} = \sum_{\kappa=1}^{K} a_\kappa \cdot \chi_{\kappa,1} \qquad (3.17)$$

Pour résoudre ces équations, on prend $\chi_{\kappa,1}$ une fenêtre dans laquelle n varie. Ensuite deux manières permettent de les construire :

- Soit on n'utilise que des y_k à l'intérieur de la fenêtre en fixant $y_k = 0$ en dehors. On obtient alors le calcul par la méthode d'autocorrélation. Dans ce cas, les coefficients d'autocorrélation $x_{k,1} = x_{1,k}$ peuvent être obtenus soit directement à partir du signal temporel, soit comme transformée de Fourier inverse du spectre de puissance obtenu, par exemple, à partir d'une simulation dans le domaine fréquentiel. La résolution des équations de Yule-Walker s'effectue en utilisant l'algorithme de Durbin [33].

- Soit on conserve les vraies valeurs de y_k à l'extérieur de la fenêtre où n varie, et on résout le système par la méthode de covariance. Pour cela il faut inverser la matrice symétrique des $x_{k,1}$.

À partir des coefficients ainsi calculés, il est possible de rechercher les racines du dénominateur de l'équation (3.11) qui correspondent aux modes de résonance du conduit vocal (formants). Or les trois premières fréquences de résonance sont suffisantes pour caractériser les voyelles du français. Cette analyse est donc utilisée pour paramétrer le signal acoustique afin d'alimenter une architecture de reconnaissance, mais aussi pour l'estimation précise des valeurs des fréquences des formants.

3.5. Les méthodes de reconnaissance automatique de la parole

Après avoir vu les trois plus populaires méthodes d'analyse du signale vocale, nous verrons maintenant deux méthodes de reconnaissance ou de comparaisons les plus utilisés actuellement dans le domaine de reconnaissance automatique qui sont : l'alignement temporelle dynamique (DTW) et les modèles de Markov cachés (HMM)

3.5.1. Alignement temporel dynamique

3.5.1.1. Idée générale

Un locuteur, même entraîné, ne peut prononcer plusieurs fois une même séquence vocale avec exactement le même rythme et la même durée. Les échelles temporelles de deux occurrences d'un même mot ne coïncident donc pas, et les formes acoustiques issues de l'étage de paramétrisation ne peuvent être simplement comparées point à point. On peut distinguer a priori deux sources de modification de l'échelle temporelle ; le changement de la vitesse d'élocution qui est représentable par une transformation linéaire de l'axe de temps, et, les variations dans le rythme de prononciation qui se traduisent par une transformation non linéaire. En fait, tout changement de vitesse d'élocution s'accompagne d'une transformation non linéaire de l'échelle temporelle, car les parties stables du signale sont plus affectés par les changements que les transitions.

Le problème de l'alignement temporel entre un mot inconnu et une référence peut être résolu de manière très efficace par un algorithme de comparaison dynamique qui va mettre en correspondance optimale les échelles temporelles des deux mots [1]. Cependant, lorsque la normalisation temporelle est basée uniquement sur un processus de comparaison dynamique une importante quantité de calculs est nécessaire pour évaluer les distances entre les événements (vecteurs de

paramètres) d'un mot inconnu et les événements de toutes les références (les nombre de distance à calculer est proportionnel au carré du nombre d'événement retenus par mots). D'autre part, les contraintes généralement employées pour optimiser les algorithmes de comparaison dynamique impose de sévères limites sur les distorsions temporelles acceptable. Il est donc intéressant d'utiliser un premier processus de normalisation temporelle agissant séparément sur chaque mot afin de simplifier la tache de l'algorithme de comparaison. Ce type de processus est basé sur le fait que les distorsions temporelles affectent surtout les zones stable du signal, et qu'un son stable peut être représenté par un nombre réduit d'événements puisqu'il est constitué d'une succession de spectre identique. Aussi, un tel processus a un double intérêt, puisqu'il réduit la quantité d'information à traiter et réalise une normalisation temporelle partielle. Sur ce principe, différents algorithmes ont été étudiés pour la reconnaissance de mots isolés et de mots enchaînés.

3.5.1.2. Principe de fonctionnement

La normalisation temporelle peut être réalisée de manière optimale au cours de la phase de comparaison en ajustant les échelles temporelles des deux mots à comparer des transformations non linéaires. Cette technique désignée par le terme de comparaison dynamique ou alignement temporel dynamique est utilisée pour la RAP depuis 1968 [1].

Si *A* et *B* sont deux images acoustique de longueur *I* et *J*, on notera $d(i, j)$ la distance entre les événements a_i et b_j. L'ajustement non linéaire de *A* et *B* est représenté par un chemin {*C(k) = (n(k),m(k)), k = 1, K*} dans *[1,I] x [1,J]* (figure 3.8). pour correspondre à une réalité physique, les fonctions *n(k)* et *m(k)* doivent respecter certaines conditions [42] :

- *Conditions de frontière : C(1)*= (1, 1) et *C(K)= (I,J)* ;
- *Continuité :* Supposons que *C(K)= (i, j)* et *C(K+1) = (i', j')* donc *i' – i ≤ 1* et *j'–j ≤ 1.* Cette contrainte est garante que *C(K+1)* doit être une cellule adjacente de *C(K)* (cela inclut la cellule adjacente diagonale) ;
- *Monotonie :* Supposons que *C(K) = (i, j)* et *C(K+1)= (i', j')* donc *i' – i ≥ 0* et *j' – j ≥ 0*. Cette contrainte force le chemin à être monotone.

Ainsi, nous supposons que les seuls chemins valides arrivant au point (i,j) viennent des point *(i-1,j), (i-1,j-1)* ou *(i,j-1)*. Cette contrainte est illustrée par la figure 3.9a (les figures 3.9b et 3.9c nous montrent d'autres solutions employées). En

supposant que les frontières des mots sont correctement définies, on prendra *C(1) = (1,1)* et *C(K) = (I,J)*.

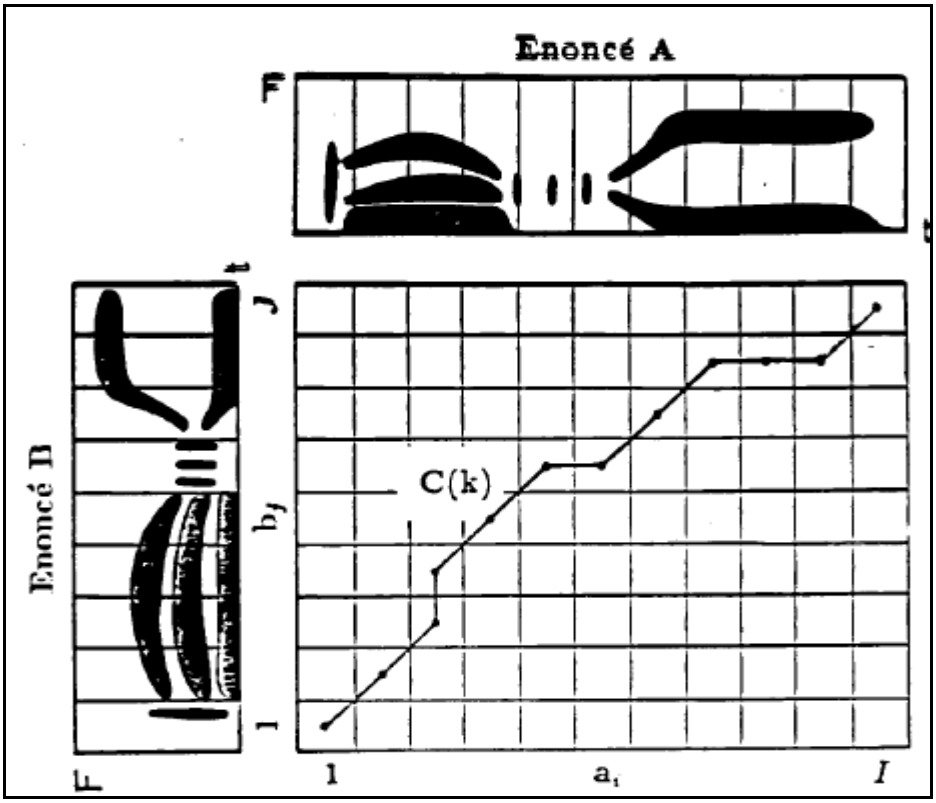


Figure 3.8 : Comparaison dynamique - Mots isolés [34]

La méthode consiste à choisir parmi tous les chemins physiquement possible, celui qui passe par les distances $d(i,j)$ les plus petites, de sorte que la somme des distances le long du chemin soit minimale, la dissemblance entre *A* et *B* étant définie comme ainsi :

$$D(A,B) = \min_{C}\left[\frac{\sum_{k=1}^{K} d(C(k)w(k))}{N(w)}\right] \qquad (3.18)$$

Où $w(k)$ est un coefficient de pondération appliqué sur le $k^{ième}$ segment du chemin *C* et $N(w)$ est un coefficient de normalisation qui dépend de la fonction w. Pour évaluer $D(A,B)$ définie par l'équation (3.18), nous devons définir fonction de pondération w, et le coefficient de normalisation $N(w)$. plusieurs types de fonction de pondération on été proposés ; en voici quatre exemples :

$$W(k) = n(k) - n(k-1) + m(k) - m(k-1) \quad (3.19)$$

$$W(k) = n(k) - n(k-1) \quad (3.20)$$

$$W(k) = m(k) - m(k-1) \quad (3.21)$$

$$W(k) = \max \{n(k) - n(k-1), m(k) - m(k-1)\} \quad (3.22)$$

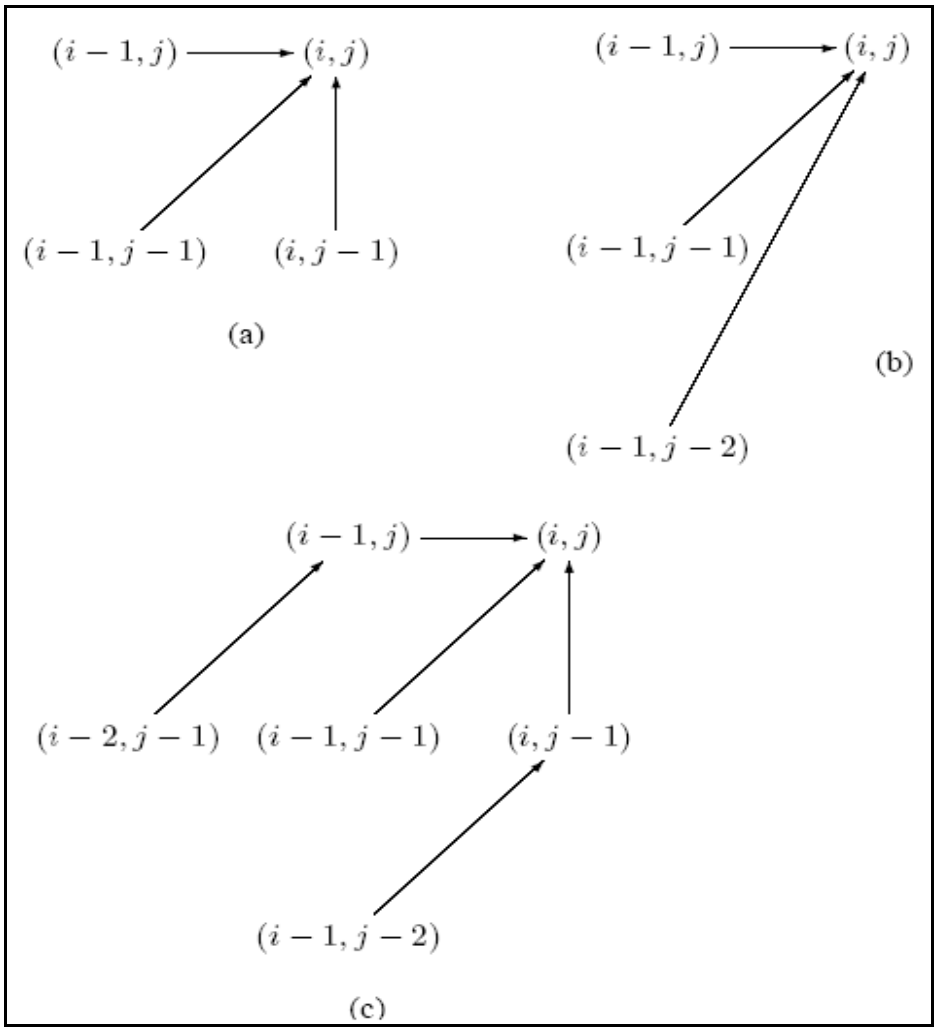


Figure 3.9 : Contraintes locales [34]

Le coefficient $N(w)$ est généralement choisi tel que $D(A,B)$ soit indépendante des longueurs de A et B, et, dans la mesure du possible pour la forme symétrique de w (équation 10).

$$N(w) = \sum_{k=1}^{K} w(k) = I + J \quad (3.23)$$

Le problème d'optimisation décrit par l'équation (3.18) peut être efficacement résolu par un algorithme de programmation dynamique. Ce dernier est grandement simplifié par le fait que $N(w)$ est indépendant de *C*, car le problème se réduit alors à la minimisation du numérateur de (3.18). Ainsi si les seules transitions autorisées sont celles indiquées sur la figure 3.9a, et si l'on retient une fonction de pondération symétrique (10) avec le coefficient de normalisation défini par (3.23), on obtient la relation récurrente locale suivante :

$$g(i,j)=\min\begin{cases} g(i-1,j)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i,j-1)+d(i,j) \end{cases} \qquad (3.24)$$

Où $g(i,j)$ est la distance cumulée le long du chemin optimal allant du point (1,1) au point (i,j). $g(i,j)$ est évaluée sur tout le domaine *[1,I] x [1,J]* qui est parcouru "colonne par colonne" ou "ligne par ligne" en partant du point (1,1). On obtient finalement :

$$D(A,B)=\frac{g(I,J)}{I+J} \qquad (3.25)$$

Nous donnons maintenant le programme C qui calcule $D(A,B)$ définie par les équations (3.24) et (3.25) :

```
# define infini 10000
g[0][0]=0;
for (j=1;j<=J;j++)     g[0][j]=infini;
for (i=1;i<=I;i++ )
{
        g[i][0]=infini;
        for (j=1;j<=J;j++)
        {
                d=dist(i,j);
                g[i][j]=min(g[i-1][j]+d,g[i-1][j-1]+d+d,g[i][j-1]+d);
        }
}
D=g[I][J]/(I+J);
```

Où $dist(i, j)$ est une fonction qui retourne la distance spectrale entre le $i^{ème}$ événement de *A* et le $j^{ème}$ événement de *B*, dans notre cas on utilise la distance euclidienne.

L'espace mémoire requis pour les $g(i, j)$ peut être considérablement réduit en remarquant que l'algorithme ne nécessite le stockage que d'un seul vecteur de dimension *J*. par contre, si l'on désire connaître le chemin optimale et non pas la distance cumulée le long de ce chemin, il est nécessaire de mémoriser le choix effectué par la fonction de minimisation pour chaque point (i,j).

3.5.2. Les modèles de Markov cachés MMC (HMM)

Nous présentons le modèle de Markov de base en premier lieu ensuite, nous verrons le modèle de Markov caché.

3.5.2.1. Modèle de Markov de base

Soit un ensemble d'instants $T= \{t_1, t2,\ldots, t_n\}$ tel que $t_1 < t_2 < \ldots < t_n$ et un ensemble d'états $S= \{S_1, S2..., S_n\}$, avec $S \subseteq E$. Un processus markovien est un processus qui satisfait la propriété suivante :

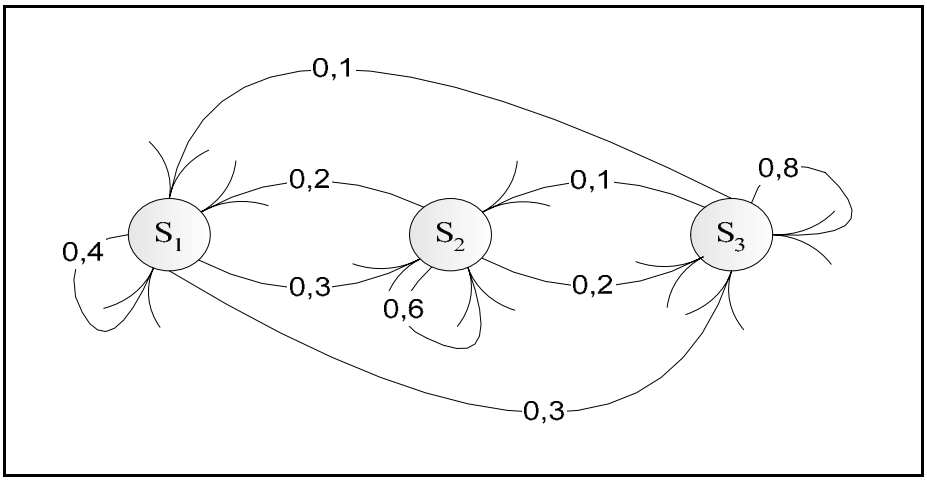
$$P\{q(t_n) \in S / q(t_{n-1}) = S_{n-1}, q(t_{n-2}) = S_{n-2}, \ldots, q(t_1) = S_1\} = P\{q(t_n) \in S / q(t_{n-1}) = S_{n-1}\}$$

Cette propriété signifie que la variable $q(t_{n-1})$ résume le passé de la séquence. Si l'espace des temps T est discrète cette propriété peut aussi s'écrire :

$$P\{q_n \in S / q_{n-1} = S_{n-1}, q_{n-2} = S_{n-2}, \ldots, q_1 = S_1\} = P\{q_n \in S / q_{n-1} = S_{n-1}\}$$

Exemple [36] :

La chaîne de Markov suivante est organisée autour des trois états qui décrivent les conditions météorologiques d'une journée donnée (pluie, nuage, soleil) :



$$S_1 = pluie$$
$$S_2 = nuage$$
$$S_3 = soleil$$

$$A = \{a_{ij}\} = \begin{bmatrix} 0,4 & 0,3 & 0,3 \\ 0,2 & 0,6 & 0,2 \\ 0,1 & 0,1 & 0,8 \end{bmatrix}$$

Supposons que le temps d'aujourd'hui (*t=1*) est ensoleillé, Qu'elle est la probabilité que le temps sera pluvieux après demain (*t=3*) ?

Il y a 3 chemins possible :

- $S_3 - S_1 \approx soleil - pluie$
- $S_1 - S_1 \approx pluie - pluie$
- $S_2 - S_1 \approx nuage - pluie$

$$p(S_3).p(S_1) = 0,8.0,1 = 0,08$$
$$p(S_1).p(S_1) = 0,1.0,4 = 0,04$$
$$p(S_2).p(S_1) = 0,1.0,2 = 0,02$$
$$P = \sum p_i = 0,08 + 0,04 + 0,02 = 0,14$$

Donc il y a 14% de chance que le temps sera pluvieux après demain.

3.5.2.2. Le modèle de Markov caché :

Une personne possédant trois pièces de monnaie les lance dans une pièce fermée. La seule chose que vous voyez est le résultat de chaque lancé *PPFPFFPP…* appelé séquence des observations. Vous ne connaissez pas l'ordre dans lequel les pièces sont lancées, non plus que les biais individuels de chaque pièce. Supposez que vous sachiez que la 3ème pièce est fortement biaisée pour F et que toutes les pièces ont une probabilité égale d'être lancée. Vous attendez alors naturellement à observer plus de *F* que de *P*. Si vous savez de plus que la probabilité de choisir la 3ème pièce (état) à partir de la première ou de la seconde (état) est nulle et que la personne était dans le premier ou le second état au début de l'observation, vous pouvez estimer que *P* et *F* apparaîtront avec la même fréquence malgré le biais. Ceci montre que la séquence des observations dépend des biais individuels, des probabilités de transition et de l'état initial [36].

Ces trois ensembles, à savoir l'ensemble des biais individuels, l'ensemble des probabilités de transition et l'ensemble des probabilités initiales des états caractérisent un *modèle de Markov caché*.

3.5.2.3. Structure d'un MMC

Un MMC est un modèle stochastique particulier [37], il représente un objet donné par deux suites de variables aléatoires. L'une dite cachée et l'autre observable. La suite cachée correspond à la suite d'états $Q_T = q_1, q_2, ..., q_T$, où les q_i puisent leur valeur parmi l'ensemble des N états du modèle. La suite observable correspond à la suite d'observations $O_T = o_1, o_2, ..., o_T$, où les o_i sont en fonctions du temps et se réalisent parmi un ensemble de M symboles observables (L. Remaki, J.G. Meumier, 2000).

Formellement, un MMC est caractérisé par ce que l'on appelle les paramètres complets du modèle $\lambda = (\pi, A, B)$, tel que :

- $\pi = \{\pi_i\}$ représente la distribution de l'état initial : $\pi_i = P(q_1 = S_i)$.
- $A = \{a_{ij}\}$ représente la distribution de probabilités sur les transitions :

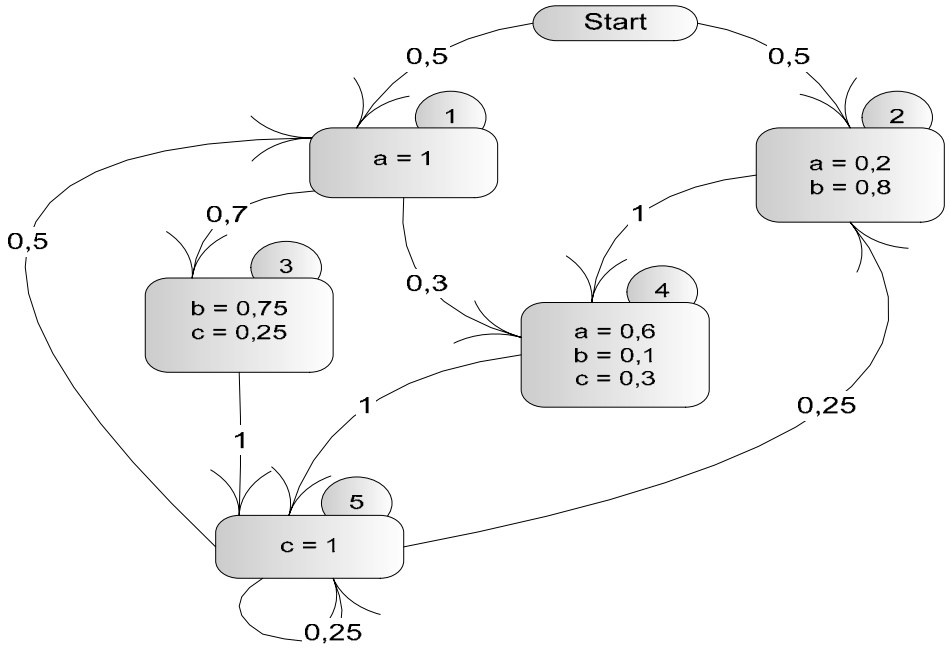
$$a_{ij} = P(q_{t+1} = S_j \,|\, q_t = S_i).$$

- $B = \{b_j(k)\}$ représente la distribution de probabilités de générations à l'état j :

$$b_j(k) = P(V_k \; aut | q_t = S_j)$$

Exemple [36] :

Soit l'HMM suivant, paramétré par *λ= (π,A,B)* :



$$\pi = \begin{bmatrix} 0{,}5 & 0{,}5 & 0 & 0 & 0 \end{bmatrix}$$

$$A = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 0 & 0{,}7 & 0{,}3 & 0 \\ 2 & 0 & 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 0 & 0 & 1 \\ 5 & 0{,}5 & 0{,}25 & 0 & 0 & 0{,}25 \end{array} \qquad B = \begin{array}{c|ccc} & a & b & c \\ \hline 1 & 1 & 0 & 0 \\ 2 & 0{,}2 & 0{,}8 & 0 \\ 3 & 0 & 0{,}75 & 0{,}25 \\ 4 & 0{,}6 & 0{,}1 & 0{,}3 \\ 5 & 0 & 0 & 1 \end{array}$$

Etant donnée les valeurs de *N, M, A, B* et *π*, l'HMM peut générer la séquence d'observation *O= o_1 o_2 … o_T* (où chaque observation o_t est l'un des symboles de *V*, et *T* est le nombre d'observations de la séquence) comme suit :

1. Choisissez un premier état $q_1=S_j$ selon le π de distribution de l'état initial.

2. Posez $t=1$.

3. Choisissez $O_1=V_k$ selon la distribution de probabilité de symbole dans l'état S_j c'est-à-dire $b_j(k)$.

4. Passage à un nouvel état $q_{i+1}=S_j$ selon la distribution de probabilité de transition d'états pour l'état S_i c'est-à-dire a_{ij}.

5. Posez $t=t+1$; si $t<T$ revenez à l'étape 3. Sinon terminez la procédure.

On appelle $\lambda=(\pi,A,B)$ les paramètres complets du modèle HMM.

3.5.2.4. Les problèmes fondamentaux des HMMs

Un HMM à trois problèmes qu'on doit résoudre pour arriver aux résultats attendus, ces derniers sont exposés dans ce qui suit.

- Le 1er Problème (évaluation) : Etant donné une séquence d'observation $O=o1\ o2\ \ldots oT$, et un modèle $\lambda=(\pi,A,B)$, le 1ier problème consiste à calculer $P(O|\lambda)$ (la probabilité de générer la séquence d'observation O suivant le modèle λ). Ce point de vue est extrêmement utile, par exemple, si nous considérons le cas dans lequel nous essayons de choisir parmi plusieurs modèles concurrents, la solution au 1ier problème nous permet de choisir le modèle qui génère le mieux les observations.
- Le 2ème Problème (décodage) : Etant donné une séquence d'observation $O=o1\ o2\ \ldots oT$, et un modèle $\lambda=(A, B, \pi)$, Qu'elle est la séquence d'états $Q=q1, q2,..., qT$ suivant le modèle λ qui a la probabilité maximale de générer O ? (le chemin le plus probable), c'est-à-dire que nous essayons de découvrir la partie cachée du modèle.

- Le 3ème Problème (apprentissage) : Etant données une séquence d'observations O, Comment peut-on re-estimer les paramètres $\lambda=(\pi,A,B)$ afin de maximiser $P(O|\lambda)$. La séquence d'observation employée pour ajuster les paramètres du modèle s'appelle une séquence d'apprentissage (ou séquence de formation puisqu'elle est employée pour former l'HMM). Le problème d'apprentissage est le problème crucial pour la plupart des applications des HMMs, puisqu'il nous permet d'adapter de façon optimale les

paramètres modèles aux données observées de formation, c'est-à-dire, pour créer les meilleurs modèles pour de vrais phénomènes.

3.5.2.5. Reconnaissance de la parole par des Modèles de Markov Cachés

Le modèle de Markov Caché est utilisé pour la modélisation acoustique dans un système de reconnaissance de la parole [28]. C'est le modèle HMM gauche-droit, ou de Bakis qui est illustré par la figure suivante:

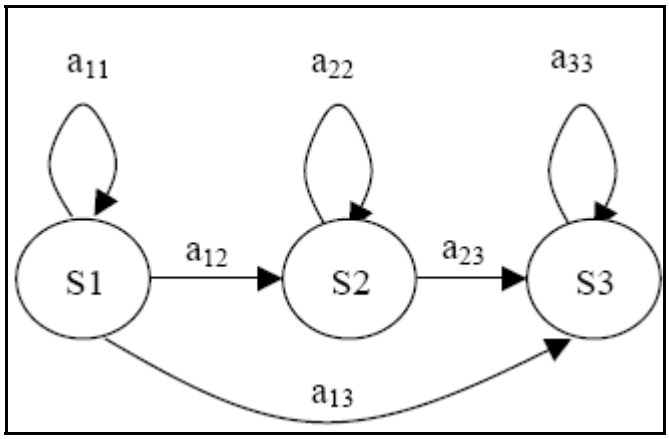


Figure 3.10 : Exemple de HMM gauche-droit

Si on veut modéliser *V* classes (phonèmes, mots ou autres unités acoustiques), chaque classe sera modélisée par un modèle de Markov Caché distinct. Donc on doit accomplir les procédures suivantes :

1) Pour chaque classe ***v***, on doit déterminer un modèle de Markov Caché $\lambda^{v}=(\pi,A,B)$ qui est optimal pour des vecteurs d'observation de la classe ***v***. C'est la phase d'apprentissage du système.

2) Pour chaque occurrence dont on veut reconnaître la classe, on détermine d'abord des vecteurs acoustiques et puis on calcule toutes les probabilités des modèles possibles $P(O|\lambda^{v})$ avec $1 < v < V$ et on choisit la classe dont le modèle donne la probabilité maximum :

$$V^{*} = \arg\max_{1\le v1\le V}\left[P(O\left|\lambda^{v}\right.)\right] \qquad (3.26)$$

On utilise l'algorithme de Viterbi pour calculer ces probabilités et la complexité est de l'ordre de $O(VxN^{2}xT)$ avec *V*, le nombre de classes, *N*, le nombre d'états et *T*, le nombre de symboles d'observations.

3.6. DTW vs. HMM

Après avoir présenté les deux méthodes DTW et HMM les plus utilisés actuellement en RAP, on peut dire que la DTW est beaucoup plus appropriée pour les SRAP de vocabulaire limité et en mode monolocuteur, tandis que les HMM ont actuellement prouvés leur puissance et robustesse pour les SRAP ayant un vocabulaire plus étendu en assurant un apprentissage robuste pour les modèles utilisés.

Néanmoins, nous pouvons résumé les différences entre les deux méthodes dans le tableau suivant :

DTW	**HMM**
Le locuteur est représenté par un ensemble de vecteurs de paramètres dans l'espace acoustique.	Le locuteur est représenté par une densité de probabilité dans l'espace acoustique.
Repose sur le principe que chaque mot est représenté par une prononciation de référence (template).	Il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle de référence du locuteur.
L'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (Time Warping).	Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'états avec des probabilités de transitions de l'un ou à l'autre, et de lois de probabilités.
Ne requiert pas une phase d'apprentissage.	Basée sur deux phases; apprentissage et décodage ou alignement
Les références sont conservées autant qu'une forme acoustique.	La référence n'est plus conservée comme une forme, mais comme un modèle (composé d'états et d'arcs et des probabilité de transition)

Table 3.1 : DTW vs HMM.

3.7. Conclusion

On peut dire que depuis une décennie, les techniques d'analyse de la parole ont connu plusieurs grandes révolutions tel que nos paroles y sont stockées sous la forme de suites de vecteurs de paramètres pertinentes, discriminantes, et robustes. Plusieurs versions d'analyse vocale ont étés apparues nous avons étudiées dans ce chapitre quelques unes d'elles (spectrogrammes, LPC, DTW, MFCC et HMM). Le chapitre suivant présentera le système de reconnaissance que nous avons développé des dix premiers chiffres de l'Arabe Standard.

CHAPITRE 4
CONCEPTION ET IMPLEMENTATION DU SYSTEME ARAD

4.1. Introduction

Après passage en revue des notions de bases nécessaires, et après avoir situé notre recherche dans un cadre théorique, Nous verrons dans ce chapitre la partie de mise en œuvre de notre système de reconnaissance automatique des chiffres arabes, ainsi que la méthodologie utilisée, qui est réalisé en deux grandes étapes : la création de la base de données sonores (élaboration du corpus), et la reconnaissance d'un signal lu.

Nous exposerons les modules de bases qui le composent depuis l'acquisition du signale de parole jusqu'aux la décision prise. Nous étalons les algorithmes implémentés pour la méthode DTW, ainsi que les différents résultats obtenus, la tâche de reconnaissance est réalisée en passant par deux grandes étapes :

- une analyse du signale de parole lu avec la méthode Cepstrale pour la génération des coefficients MFCC;
- application de la méthode DTW pour la comparaison du signal avec la base sonore.

4.2. Spécification des besoins

Pour assurer un bon développement de notre démarche et pour obtenir une meilleure organisation de notre travail, nous avons défini les objectifs visés. Ces derniers peuvent se résumer par les points suivants :

- notre principal objectif est de réaliser un système de reconnaissance automatique des chiffres prononcés en Arabe Standard ;

- assurer une interactivité et une simplicité d'utilisation, puisque notre travail est destiné à des utilisateurs de plusieurs disciplines (phonéticiens, électroniciens, informaticiens…) ;

4.2.1. Les cas d'utilisation (use cases)

Dans cette phase on doit représenter le comportement et la réaction de notre outil face aux exigences et aux actions des différents acteurs (tout élément qui interagit avec notre système), dans notre cas ces derniers peuvent se résumer par *un acteur principal* qui est l'utilisateur de notre système.

Pour bien modéliser, et pour bien détailler les fonctionnalités de notre système, on a utilisé des schémas proches des notations d'UML (Unified Modeling Langage).

4.2.2. Diagrammes de cas d'utilisation

Chaque acteur de système peut réagir d'une ou de plusieurs manières en réponse aux différentes interactions avec le système. On peut lui associer une ou plusieurs fonctionnalités (Tableau 4.1).

Acteur	Buts utilisateur
Utilisateur	Ouvrir le système Entrer le signal vocal (ouvrir, ou enregistrer) Analyser le signal Sauvegarder le résultat de l'analyse Lancer la reconnaissance Consulter les résultats Générer acoustiquement le mot reconnu Quitter l'application.

Table 4.1 : Cas d'utilisations de notre système ARAD.

Il est clair que notre système de reconnaissance automatique de chiffres en Arabe Standard (ARAD) doit comprendre trois modules :

- la création de la base des sons (corpus) ;

- l'analyse Cepstrale (coefficients MFCC) du signal lu;

- la reconnaissance vocale de ce signal.

Cela donne naissance aux cas d'utilisation suivants :

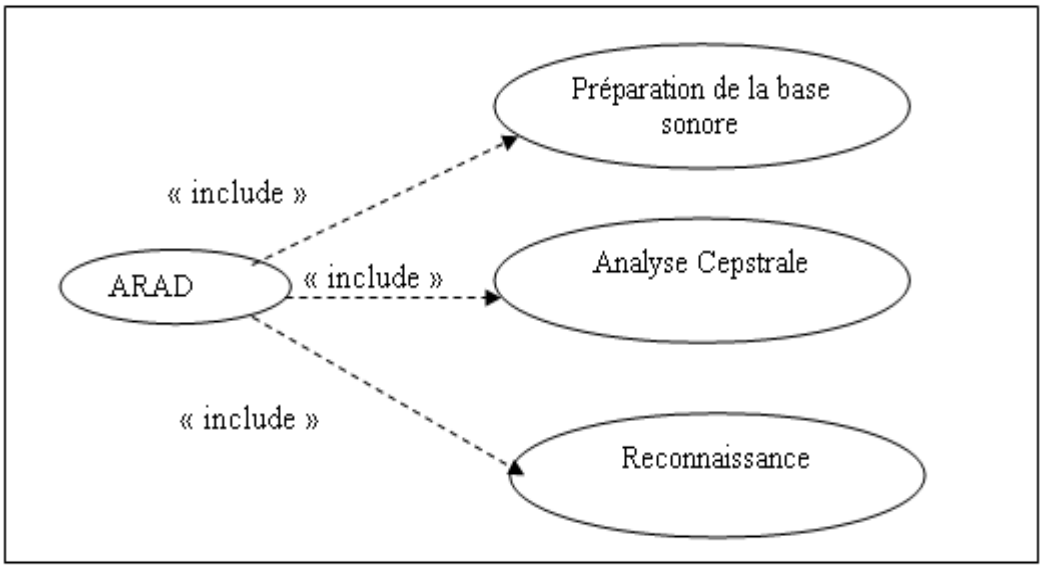


Figure 4.1 : Diagramme des cas d'utilisation.

- Cas d'utilisation « préparation de la base sonore »

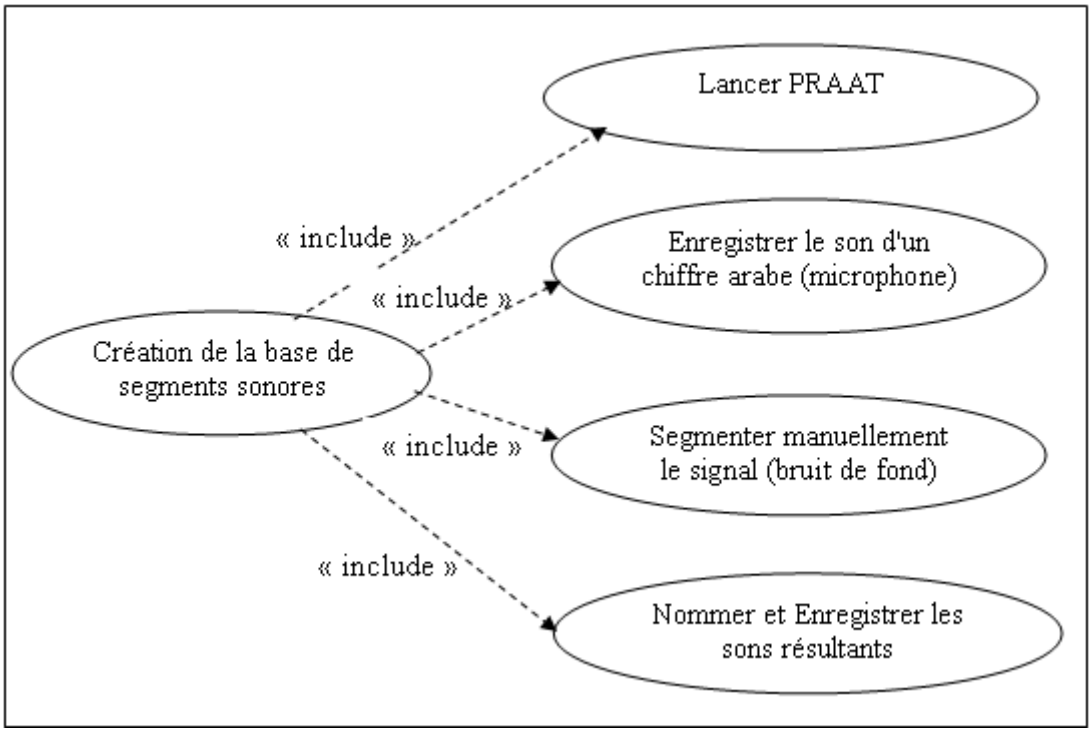


Figure 4.2 : Diagramme de use cases de cas " préparation de la base sonore".

- Cas d'utilisation « Analyse Cepstrale »

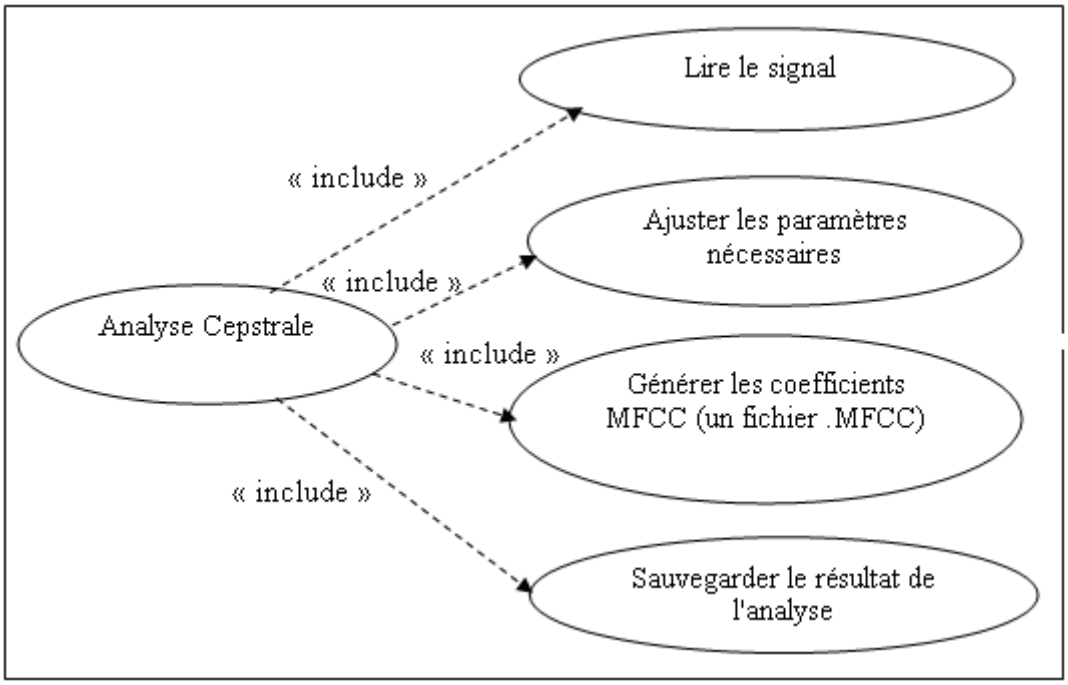


Figure 4.3 : Diagramme de use cases de cas "Analyse Cepstrale".

- Cas d'utilisation « Reconnaissance»

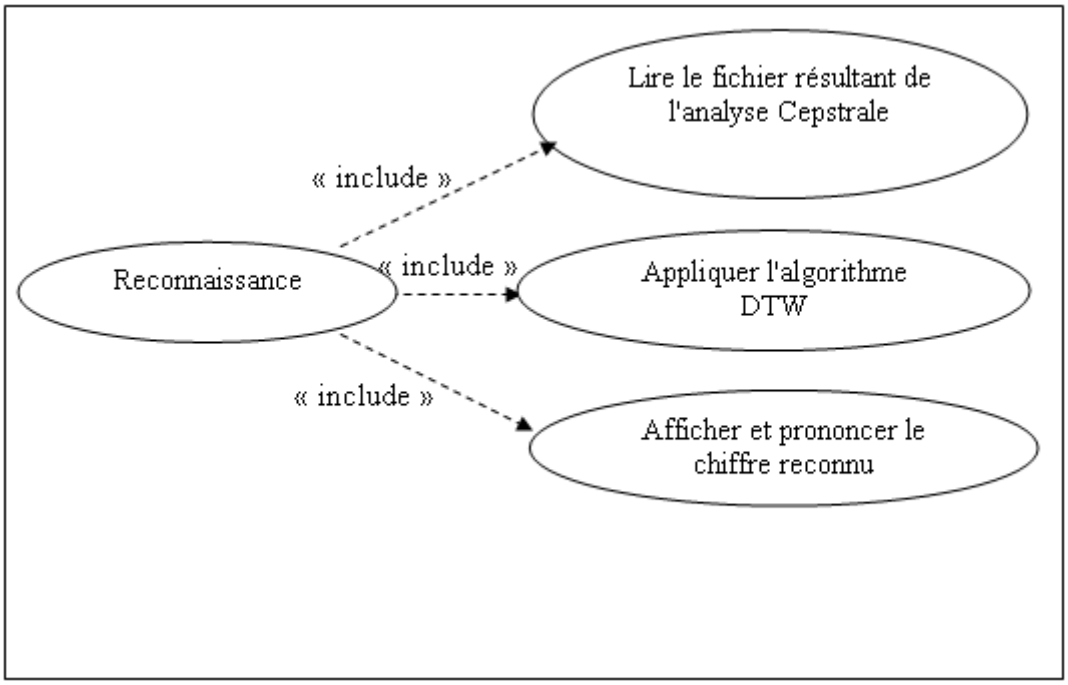


Figure 4.4 : Diagramme de use cases de cas "Reconnaissance".

4.3. But de notre Travail

En reconnaissance de la parole, l'étape d'extraction des caractéristiques, appelée communément l'étape d'analyse, peut-être réalisée de plusieurs manières. En effet, les vecteurs acoustique sont généralement extraits à l'aide de méthodes temporelles comme le codage linéaire prédictif (LPC) ou de méthodes Cepstrales comme le codage MFCC, ainsi que le codage PLP (Perceptual Linear Predictive coding) qui est un exemple de l'application des connaissances du système auditif humain en reconnaissance de la parole. L'extraction de caractéristiques est un élément clé pour la mise au point d'un système de reconnaissance. De nombreux travaux ont montré l'importance de cette étape.

Notre Objectif ici est d'élaborer un système de reconnaissance automatique des dix premiers chiffres arabes en mode mono-locuteur. Nous nous intéressons exclusivement à l'étape d'analyse du signale de parole qui permet d'extraire les vecteurs acoustiques caractérisant ce dernier. Cette étape est très importante et primordiale dans le processus de reconnaissance automatique, puisque elle produit en sortie un ensemble de paramètres jugés pertinents et suffisants pour la bonne exploitation du signal de parole, sur ce même ensemble que nous allons appliquer les algorithmes de reconnaissances et de comparaison.

Nous devrons faire face à deux problèmes, d'une part, le choix de la technique d'analyse utilisée, et d'autre part, le choix des paramètres et leurs nombres.

4.4. Structure du programme

Pour faciliter l'implémentation et la modification de notre programme ARAD (Automatic Recognition of Arabic Digit), nous avons utilisé l'approche modulaire, ce concept rend le programme compréhensible d'une part et diminue le coût de développement de chaque module d'une autre part. Nous avons utilisé aussi les notions de la programmation orientée objet qui se combine particulièrement avec les techniques modulaires. Nous devons donc

distinguer différents modules qui structurent le programme, ceci est schématisé comme suit : (Figure 4.5)

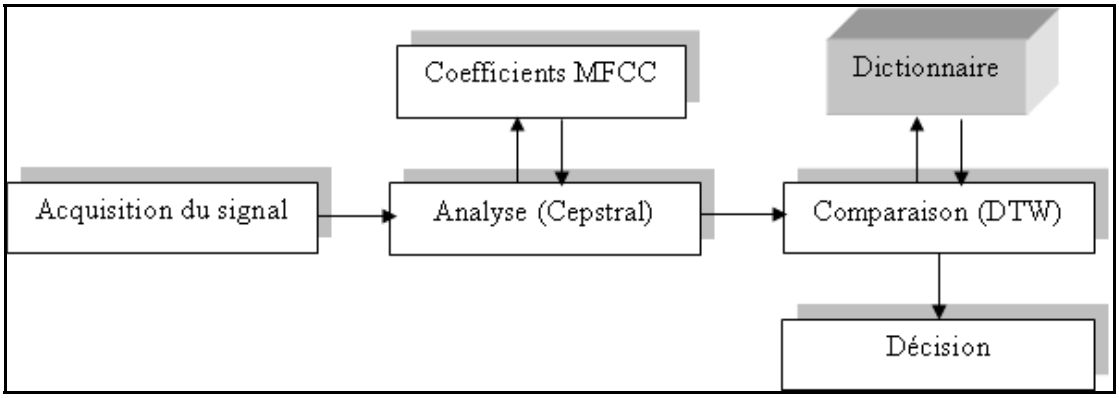


Figure 4.5 : Architecture de notre système ARAD

L'objectif consiste à détailler le rôle de chaque module, en expliquant l'intérêt des liens qui assurent la coopération entre eux.

4.4.1. Acquisition du signal

Ce module réalise l'acquisition de l'onde acoustique de parole captée par un microphone et la convertit en une forme exploitable par la machine. Il existe de nombreux types de microphones mais tous assurent la même fonction : transformer les fluctuations de pression causées par l'onde acoustique de parole en un signal électrique. Ce signal subit ensuite une conversion analogique-numérique, c'est-à-dire qu'il est discrétisé à la fois en temps (échantillonnage) et en valeur (quantification).

On obtient ainsi un signal numérique sous la forme d'une séquence d'échantillons qui mesurent l'amplitude du signal du microphone à des instants régulièrement espacés, et l'amplitude de chaque échantillon est représentée sous sa forme numérique, utilisable par la machine. Le choix de la fréquence d'échantillonnage est généralement fonction de l'application visée et de la plate-forme utilisée [38].

4.4.1.1. Fréquence d'échantillonnage

Certaines réflexions sur la fréquence d'échantillonnage sont nécessaires au préalable. D'après le théorème de Shannon: "l'information véhiculée par un signal dont le spectre est à support borné n'est pas modifiée par l'opération d'échantillonnage à condition que la fréquence d'échantillonnage soit au moins deux fois plus grande que la plus grande fréquence contenue dans le signal" [28].

L'information acoustique pertinente du signal de parole se situe principalement dans la bande passante [50 Hz - 8 kHz], la fréquence d'échantillonnage devrait donc au moins être égale à 16 kHz, selon le théorème de Shannon. Pour le cas de notre application, on a utilisé une fréquence d'échantillonnage de l'ordre de 22050, la valeur prise par défaut par le logiciel *PRAAT* (*PRAAT* est un logiciel d'analyse et de transcription phonétique, comportant des fonctionnalités importantes pour l'enregistrement, pour la manipulation et pour la synthèse de parole [43]).

4.4.1.2. Elaboration du corpus

La plupart des travaux effectuer dans le domaine de la communication parlée nécessite souvent l'enregistrement, et la manipulation de corpus de parole continue, et cela pour mener à bien les études sur les effets contextuels, sur les indices phonétiques, et sur les variabilités intra et inter locuteurs.

Nous avons enregistré trois corpus contenant chacun dix sons des dix premier chiffres de l'Arabe Standards ([wahid], [ithnané], [thalatha], [arbaa], [khamsa], [sita], [sabaa], [thamania], [tisaa], [aachara]) dans un environnement bruité et nous avons changé la vitesse d'élocution d'un corpus à l'autre sans changer de locuteur. L'étape d'analyse peut donc commencer.

4.4.2. L'analyse Cepstrale (MFCC)

L'analyse acoustique traite le signal vocal et en extrait les vecteurs acoustiques qui seront utilisés pour la phrase de reconnaissance suivante. Dans cette étape le signal de parole est transformé en une séquence de vecteurs acoustiques pour diminuer la redondance et la quantité de données à traiter. Puis une analyse spectrale par la transformation de Fourier Discrète,

est effectuée sur une trame de signal (généralement de taille 20 ou 30 ms). Dans cette trame, le signal vocal est considéré comme suffisamment stable et on en extrait un vecteur de paramètres jugé suffisant pour la bonne exploitation du signal vocal, dans notre travail nous avons choisi d'utiliser des coefficients MFCCs issus d'une analyse cepstrale du signal en question [28].

La méthode d'extraction des MFCCs, la méthode cepstrale, est l'une des méthodes de calcul des vecteurs acoustiques les plus populaires dans le domaine de Reconnaissance Automatique de la Parole. Nous avons également décidé de l'utiliser dans notre contexte d'applications et nous avons choisi un ensemble de 12 coefficients.

Nous exposons maintenant comment nous avons réalisé l'analyse Cepstrale à l'aide d'un outil d'analyse de parole, *PRAAT*, et nous montrons les différents paramètres requis pour l'analyse que nous avons choisie, et enfin l'exploitation des coefficients MFCC résultants.

Etape1 : Lecture du fichier à analyser et le choix de la méthode MFCC

- Lancer *PRAAT*
- Ouvrir le fichier son :
    - *Read > Read from file* → pour ouvrir un fichier son
    - *Edit* → pour le visionner
    - *File > Extract Selection* → pour "découper" le son
    - *Write > Write to .wav file* → pour sauvegarder un fichier son
- Sélectionner le fichier à analyser
- Choisir la méthode Cepstrale :
    - *Formants & LPC > To MFCC*

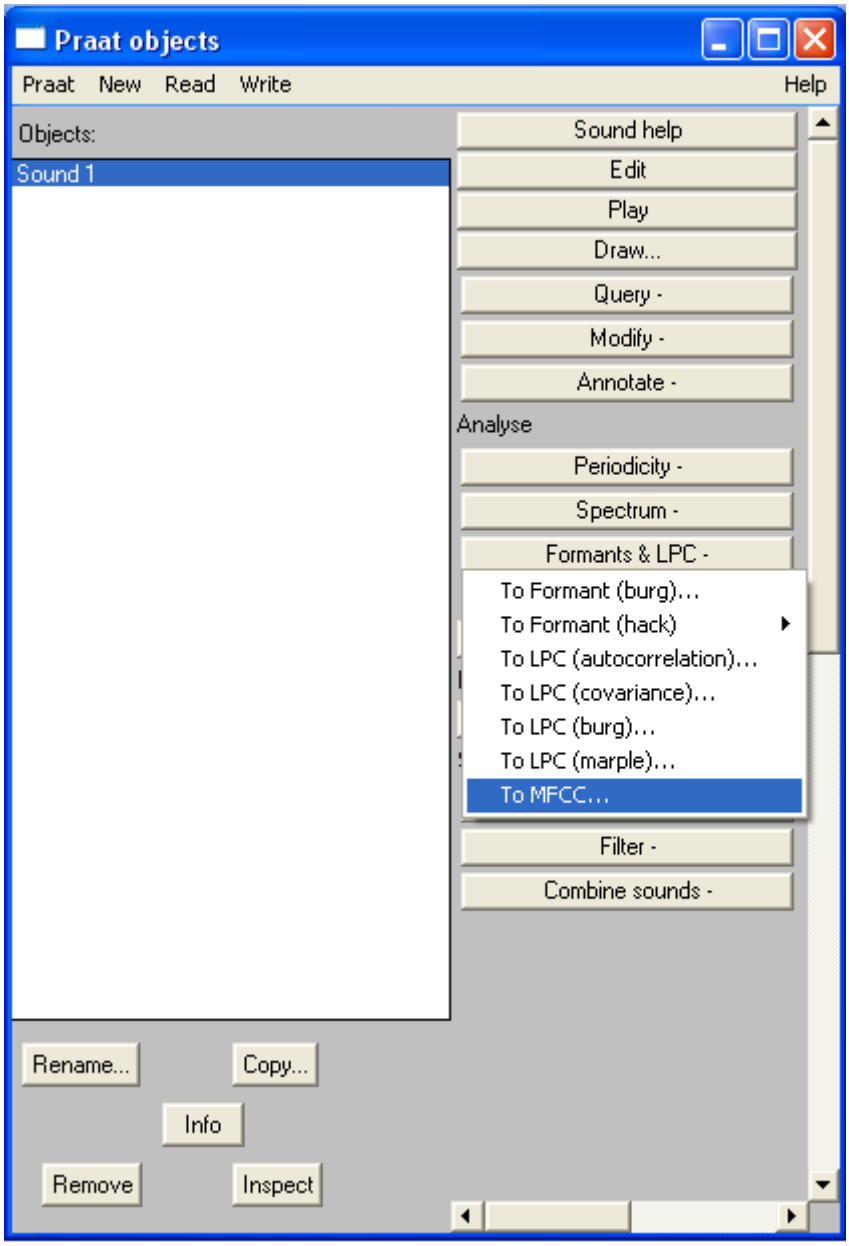


Figure 4.6 : Lecture du fichier à analyser et le choix de la méthode MFCC.

Etape 2 : détermination des paramètres requis pour l'analyse

- Nombre de coefficients : 12 ;
- Durée des fenêtres : 30 ms ;
- Durée entre les fenêtres : 10 ms.

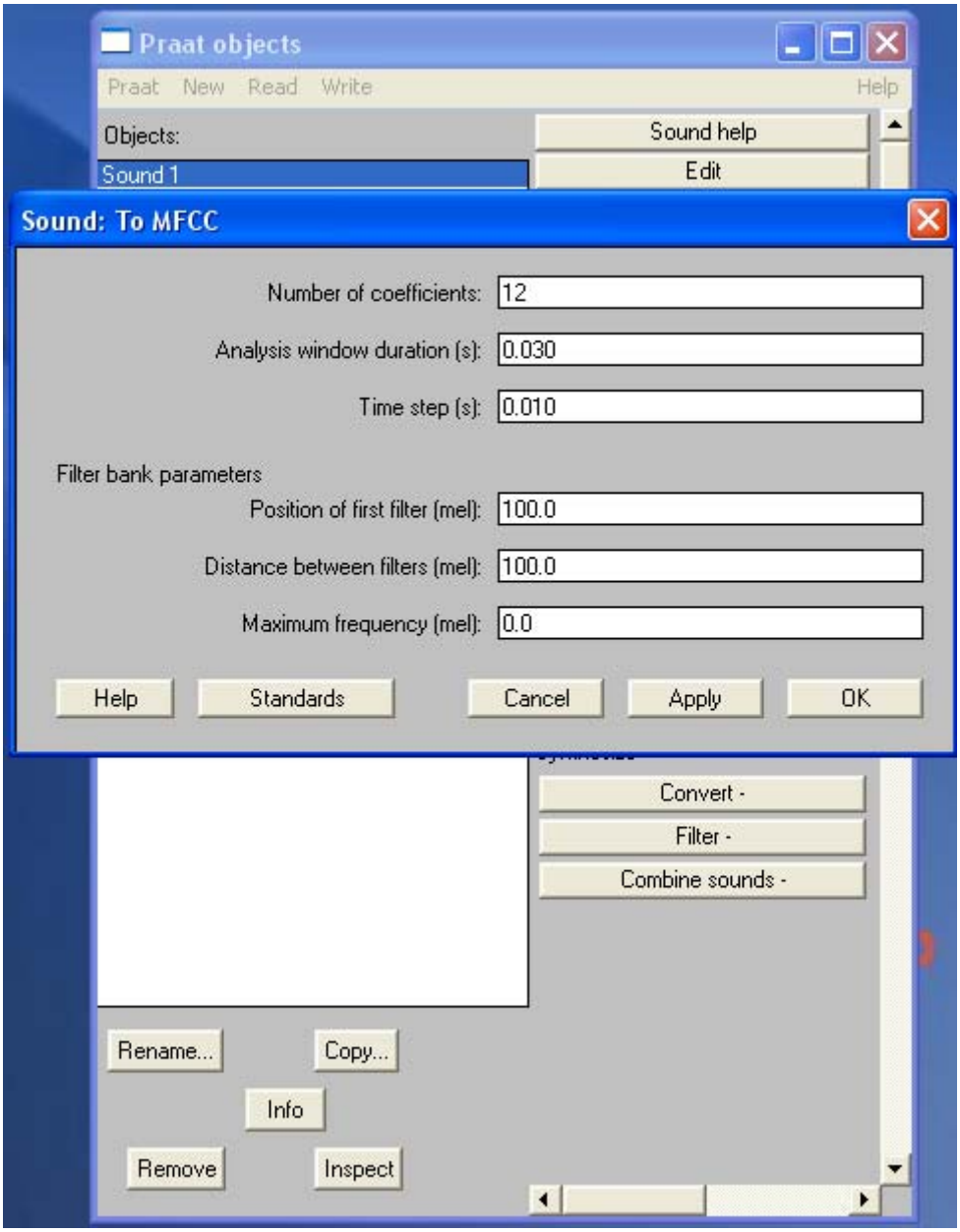


Figure 4.7 : Choix des paramètres pour lancer la MFCC.

Etape 3 : Résultats de l'analyse

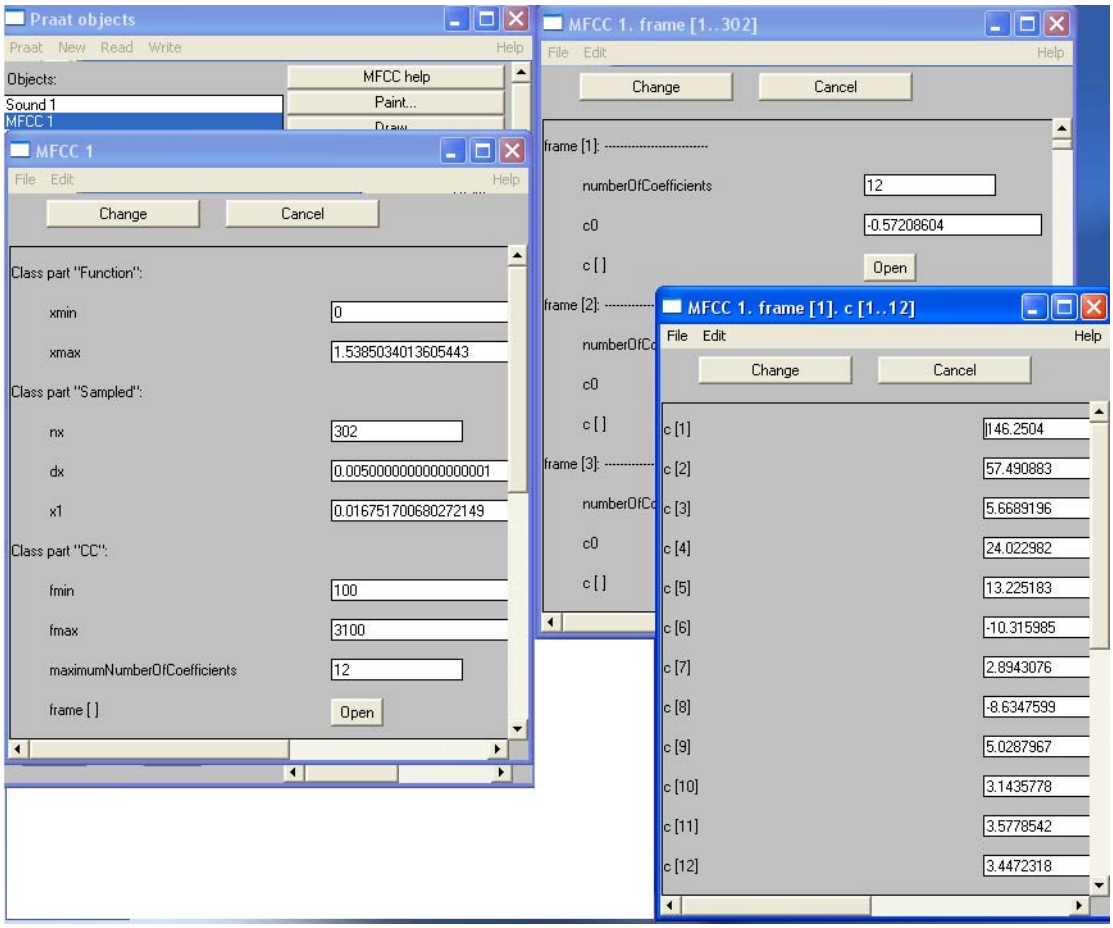


Figure 4.8 : Résultats de l'analyse (les 12 coefficients MFCC).

Il ne reste maintenant qu'à enregistrer les résultats dans un fichier format texte portant l'extension .MFCC (*Write > Write to txt file*), pour être utilisé dans l'étape qui suit.

4.4.3. La méthode de comparaison (DTW)

Notre système de reconnaissance de la parole est basé sur l'algorithme de DTW, il essaie d'évaluer la distance entre une observation et une liste de références (dictionnaire), ainsi la référence pour laquelle cette distance est minimale permet de dire de quel mot il s'agit.

L'évaluation de la distance entre deux signaux ne s'effectue pas avec les signaux eux-mêmes. Cela ferait beaucoup trop de calculs. Il s'agit donc dans un premier temps de trouver une meilleure représentation des signaux. C'est ici qu'intervient l'étape d'analyse; les coefficients MFCC.

Nous avons programmé la méthode de DTW en utilisant, pour la comparaison, des coefficients MFCC. La partie apprentissage concerne l'enregistrement des corpus de sons afin de concevoir le dictionnaire avec lequel sera comparé les signaux de test.

Des problèmes de reconnaissance peuvent apparaître selon les conditions dans lesquelles le signal à tester est enregistré. Si le mot est prononcé plus ou moins proche du microphone les taux de reconnaissance peuvent varier grandement. Cependant si l'utilisateur prononce le mot toujours à la même distance et avec la même intensité, les taux de reconnaissance sont très satisfaisants.

Il résulte néanmoins que la représentation à l'aide des coefficients MFCC fournit de meilleurs résultats, et supporte mieux les limitations exposées liées au problème de la capture du signal.

L'algorithme principal de la DTW comporte trois étapes suivantes :

- l'acquisition du fichier du son à tester ;
- l'extraction des coefficients MFCC ;
- la comparaison avec le dictionnaire des références.

4.4.4. La décision

Ce dernier module de notre application représente l'interface avec laquelle l'utilisateur interagit avec le système. Après que l'utilisateur a entré son signal vocal, il lance la recherche et attend les résultats. Le programme affiche à la fin le chiffre reconnu tout en le prononçant, comme sera évoqué ultérieurement.

4.5. Présentation du notre logiciel ARAD

Les systèmes de Reconnaissance Automatique de l'Arabe Standard représentent un domaine de recherche très actif mais difficile à mettre en œuvre. Afin de faciliter cette tâche nous avons proposé un outil; l'exécution de ce dernier commence par le déclenchement d'une interface principale (Figure 4.9).

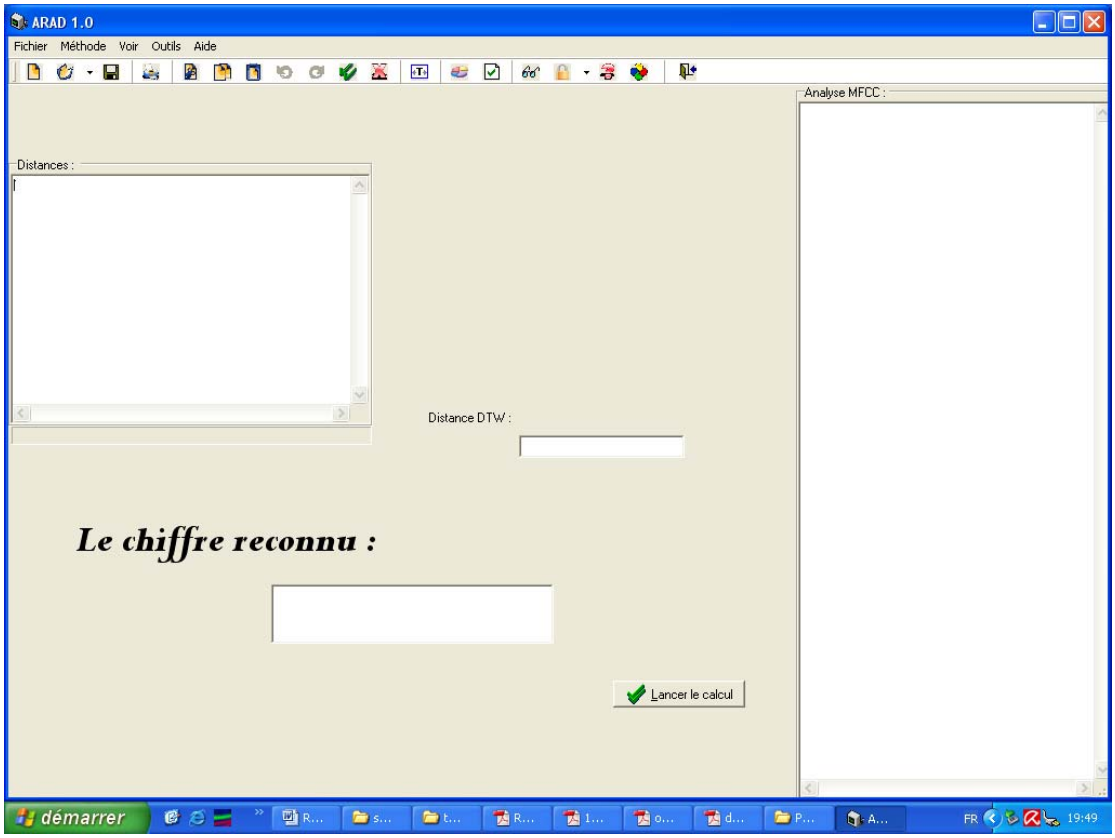


Figure 4.9 : Fenêtre principale de ARAD.

ARAD comporte les composants suivants:

- Un menu principal ;
- Une barre d'outil ;
- Une zone affichant les distances calculées en temps réel ;
- Une zone affichant les coefficients MFCC du signal entré ;
- Une zone affichant la distance DTW minimale retrouvée ;
- Une zone affichant le chiffre reconnu ;
- Un bouton pour lancer le calcul ;

A la fin du calcul l'utilisateur peut entendre le système prononçant le chiffre reconnu.

4.6. Tests et résultats

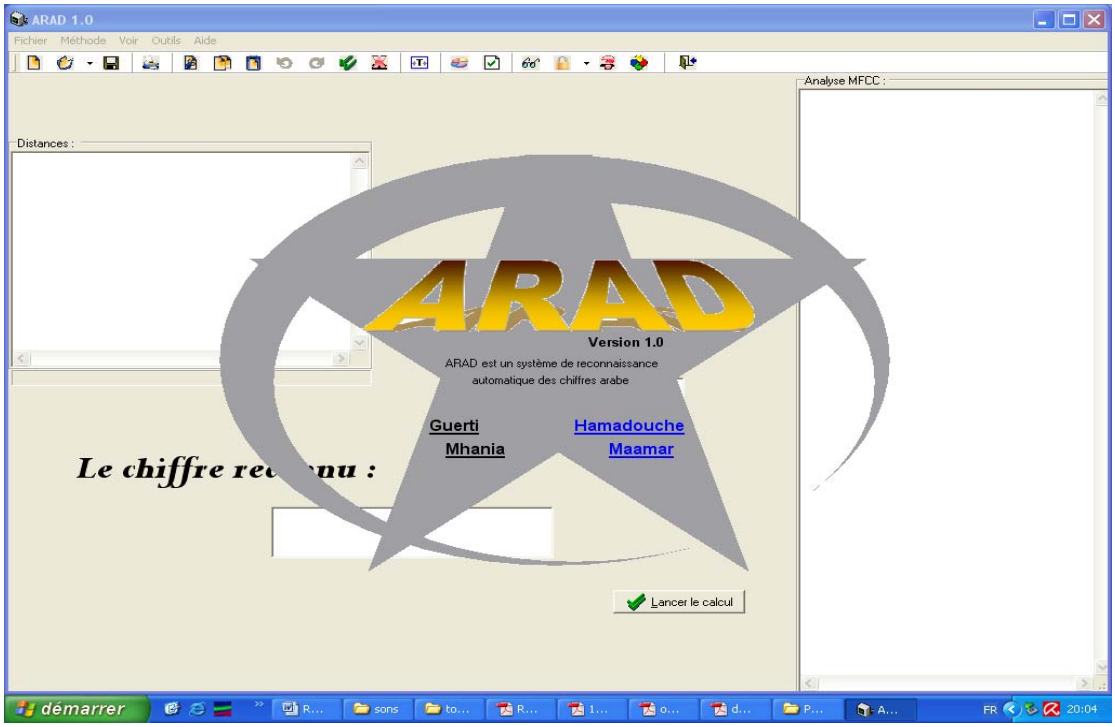


Figure 4.10 : Système ARAD

Nous avons appliqué la reconnaissance sur un corpus contennant des chiffres arabes de un à dix prononcés par un locuteur de sexe masculin en langue arabe standard. Les résultats obtenus sont résumés dans les exemples suivants :

Exemple 1 : Reconnaissance du chiffre "WAHID"

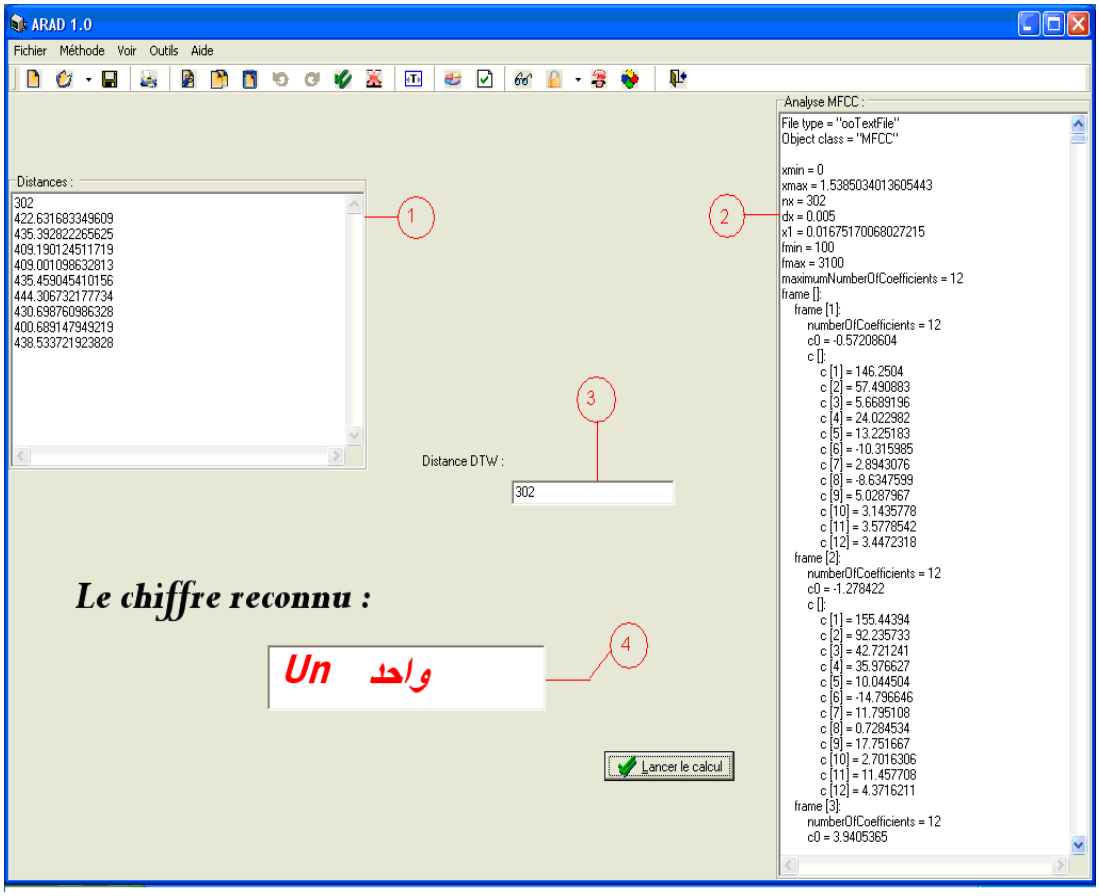


Figure 4.11 : Reconnaissance du chiffre /un/ [wahid].

1) les distances calculées par la méthode DTW
2) les coefficients calculées par la MFCC
3) la distance cumulée résultante à partir de la DTW
4) le résultat de la reconnaissance

Exemple 2 : Reconnaissance du chiffre [AACHARA]

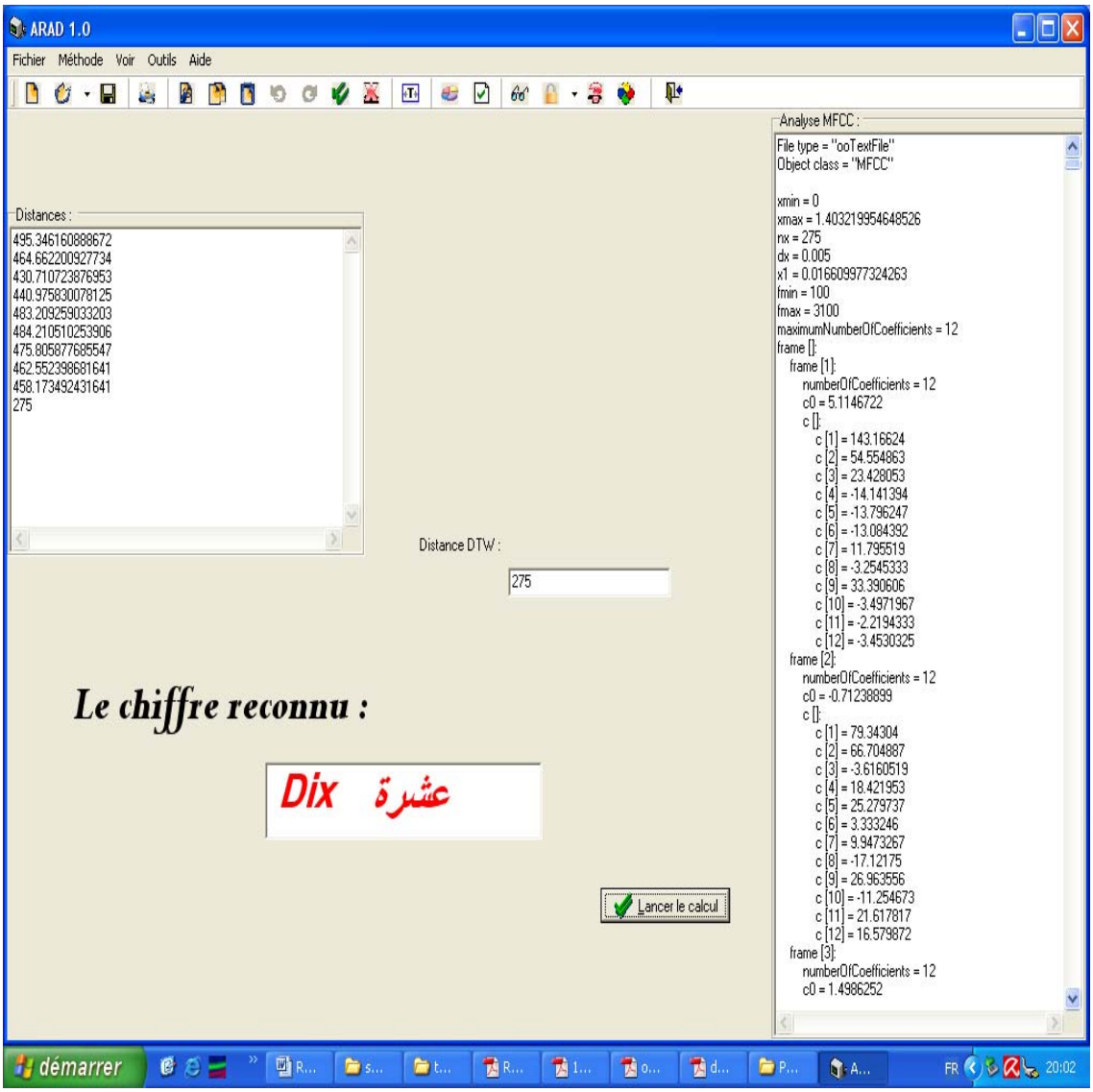


Figure 4.12 : Reconnaissance du chiffre /dix/ [AACHARA].

4.7. Conclusions

Dans ce chapitre nous avons expliqué la partie de conception et de mise en œuvre, dans laquelle nous avons présenté notre système de reconnaissance des dix premiers chiffres de l'Arabe Standard (ARAD) qui est basé dans sa partie reconnaissance sur la méthode DTW, et dans sa partie analyse sur la Cepstrale pour en extraire les coefficients MFCC, nous avons vu que le système présente un taux de reconnaissance de 96% selon les tests que nous avons effectués, et permet effectivement de faire la reconnaissance de mots isolés en mode mono locuteur des dix premiers chiffres de l'Arabe Standard.

CONCLUSION

Nous nous étions fixée plusieurs objectifs pour cette recherche : celle de découvrir les traits définitoires de la voix humaine, de décrire les différents étapes et organes de production de la voix et de décortiquer un système de reconnaissance automatique de la parole en ces principales étages.

Nous avons à cette fin, délimité dans une première étape des notions fondamentale concernant le signale de parole et ces différents paramètres pertinents, depuis sa production par l'appareil phonatoire humain jusqu'à ça perception par l'oreille, ainsi qu'une vue d'ensemble sur la langue arabe et ses particularités.

Nous avons présenté en suite un état de l'art du domaine de la reconnaissance automatique de la parole. Après en avoir posé les grands principes, nous avons dressé un bref historique des travaux menés depuis plus d'un demi-siècle. Ce domaine se caractérise actuellement à la fois par le développement d'applications pratiques dans des secteurs variés et par un effort de recherche toujours important, notamment pour augmenter la fiabilité et la robustesse des systèmes dans le cadre de la communication Homme-Machine.

Dans une deuxième partie nous avons abordé les techniques d'analyse acoustiques les plus couramment utilisés en vue de reconnaître la parole ; basés sur la méthode cepstrale ou la prédiction linéaire, ils sont utilisés pour extraire des indices permettant de classifier les différents sons de parole ou comme un simple moyen de représenter de manière concise l'information pertinente pour la RAP.

Nous avons décrit quelques architectures et les différents principes appliqués aux problèmes spécifiques de la reconnaissance de la parole. Ainsi, le principe de la programmation dynamique utilisé dans les premiers systèmes de reconnaissance permet de prendre en compte les différences de durée lors de diverses prononciations d'un mot. Il a été repris dans

l'algorithme de Viterbi pour les modèles de Markov. Mais ces derniers ont l'avantage de disposer d'un algorithme d'apprentissage qui permet d'ajuster automatiquement les paramètres du modèle à partir d'exemples. D'un autre côté, l'approche analytique consistant à rechercher des invariants dans le signal acoustique puis à modéliser l'expertise humaine a beaucoup contribué à une meilleure compréhension des problèmes de perception et compréhension de la parole. Ces architectures pourront être utilisées avec des paramètres articulatoires issus d'une inversion au lieu d'utiliser directement le signal acoustique.

Dans notre travail, nous avons présenté un système de reconnaissance automatique des dix premiers chiffres de l'arabe standard ainsi que les résultats obtenus. Il a permet d'obtenir pour les mots isolés et en l'absence de bruit des taux de reconnaissance tout à fait honorables et acceptable.

Au terme de ce bilan rapide sur la reconnaissance vocale, on a pu constater que ce domaine est particulièrement vaste et qu'il n'existe pas de produit miracle capable de répondre à toutes les applications. Le bruit, par exemple, non traité par ce document, reste un frein à la généralisation des systèmes de reconnaissance.

La reconnaissance vocale reste un compromis entre la taille du vocabulaire, ses possibilités multilocuteur, son encombrement physique, sa rapidité, temps d'apprentissage, etc.

La puissance des outils de calcul actuels et les capacités d'intégration des systèmes ont provoqué un regain d'intérêt depuis ces dernières années chez les industriels. En effet, ces derniers voient dans la reconnaissance vocale, " le plus commercial ", permettant de faire la différence avec la concurrence.

Un rapide tour d'horizon sur les très nombreuses publications permet de se fixer les idées sur la nature des travaux en cours. Hormis les produits dédiés à la reconnaissance de la voix, les systèmes à approche analytique

(HMM) donnent aujourd'hui les meilleurs résultats et ont actuellement le vent en poupe.

Quant aux perspectives d'avenir, l'optimisme est plus mesuré que dans le passé. Sans risque, on peut affirmer que le problème général du traitement automatique de la parole ne sera sans doute pas réglé avant le milieu du siècle prochain. Dans une prochaine étape de notre travail, nous envisagent faire la reconnaissance de touts les chiffres de l'AS en mode multi locuteurs de la parole continue; d'autres méthodes d'analyses et de reconnaissances pouvant aussi êtres développer, nous donnons l'exemples des réseaux de neurones et les méthodes hybrides.

APPENDICE

LISTE DES SYMBOLES ET DES ABREVIATIONS

ARAD : Automatic Recognition of Arabic Digit.

AS : Arabe Standrad

BE : Bande Étroite.

C : Consonne

C-STAR : Consortium for Speech Translation Advanced Research.

DTW : Dynamic Time Warping.

dB : deciBel

DAP : Décodage Acoustico-Phonétique.

FFT : Fast Fourier Transform.

GDW : Gaussian Dynamic Warping.

HMM : Hidden Markov Model.

Hz : Hertz

IFFT : Inverse Fast Fourier Transform.

LPC : Linear Predictive Coding.

LB : Large Bande.

MMC : Modèle de Markov Caché

MFCC : Mel-Frequency Cepstrum Coefficients.

PLP : Perceptually-based Linear Prediction.

RAL : Reconnaissance Automatique du Locuteur.

RAP : Reconnaissance Automatique de la Parole.

RF : Reconnaissance de Formes.

SRAP : Système de Reconnaissance Automatique de la Parole.

TAP : Traitement Automatique de la Parole.

UML : Unified Modeling Langage.

V : Voyelle

REFERENCES

1. Calliope, "La parole et son traitement automatique", Edition Masson, Paris, France, 1989.

2. Dutoit . T, "Je parle, donc je suis ? un bilan des développements récents en traitement automatique de la parole ", Revue de la Société des Arts, Faculté Polytechnique de Mons, Belgique, 2001.

3. Koreman. J, Andreeva. B et Strik. H, "Acoustic parameters versus Phonetic Features in ASR", Dans International Congress of Phonetic Sciences, pp. 549-553. San Francisco, USA, août 1999.

4. .Dutoit. T, "Introduction au Traitement Automatique de la Parole", Faculté Polytechnique de Mons, Belgique, Première édition, 2000.

5. Drygajlo. A, "traitement de la parole", Groupe de Traitement de la Parole et de Biométrie (GTPB), Institut de Traitement des Signaux (ITS), section d'électricité (SE), EPFL, IDIAP, Martigny Lausanne 2003.

6. Meuwly. D, " Reconnaissance du locuteur en sciences forensiques: L'apport "d'une approche automatique", Institut de Police Scientifique et de Criminologie, Université de Lausanne, Lausanne, 6 décembre 1996.

7. Lassagne. F, "Maîtriser la voix, de la pédagogie à la Synthèse artificielle : Un phénomène sonore aux frontières des Sciences", Université Paris 7, Denis Diderot, Dess communication et information Scientifiques, Année universitaire 2001 – 2002.

8. Zimmermann. P, "Analyse de l'Influence des Conditions d'Enregistrement dans la Reconnaissance Automatique de Locuteurs en Sciences Forensiques", Séminaire de Sciences Forensiques, 10 juin 2005.

9. Hennebert. J, "Traitement de la Parole, Cours 2: Signal de parole, Production – Perception – Analyse", University of Fribourg, 27 mars 2006.

10. Drygajlo. A, "Sécurité multimédia, Chapitre 10 : Reconnaissance vocale et sécurité", http://scgwww.epfl.ch/courses/Traitement_de_la_parole-2005-2006-pdf/Drygajlo-Reconnaissance-du-locuteur.pdf

11. Droua-Hamdani. G, "prédiction de la durée segmentale des phonèmes de l'Arabe Standard", 2003.

12. "la voix humaine", du site: http://www.musimem.com/voix-humaine.htm.

13. Chetouani. M, Gas. B et Zarader. J.L, "Coopération entre codeurs neuro-prédictifs pour l'extraction de caractéristiques en reconnaissance de phonèmes", Laboratoire des Instruments et Systèmes d'Ile-De-France, Université Paris VI, France, 2004.

14. Rachedi. J, "Reconnaissance et classification de phonèmes", Mémoire de Master en Sciences et Technologie de l' UPMC, Spécialité SAR, Parcours ATIAM, Laboratoire IRCAM, Paris, France, Mars / Août 2005

15. Barras. C, "Classification et reconnaissance de la Parole", LIMSI-CNRS, séminaires du cycle classification 2003 - 2004.

16. Bezat. M.C, "Qualification acoustique du typage sonore", Application au typage sport, Rapport de stage de DEA ATIAM, (Paris VI, ENST, Aix-Marseille II, UJF Grenoble I), Mars-août 2003.

17. Al-Zabibi. M, "An Acoustic–Phonetic Approach in Automatic Arabic Speech Recognition", The British Library in Association with UMI, 1990.

18. Douzidia. F.S, "Résumé automatique de texte arabe", Université de Montréal, Département d'informatique et de recherche opérationnelle, Faculté des arts et des sciences, Septembre 2004.

19. Muhammad. A, "Alaswaat Alaghawaiyah", Daar Alfalah, Jordan, 1990.

20. Elshafei. M, "Toward an arabic Text-To-Speech system," The Arabian J. Science and Engineering vol. 4B N° 16, pp. 565–583, 1991.

21. Baloul. S, "Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé," Thèse de Doctorat, Université du Maine, Le Mans, France, 2003.

22. Satori. H,Harti. M et.Chenfour N, "Système de reconnaissance automatique de l'arabe basé sur CMUSphinx", http://arxiv.org/pdf/0704.2201.pdf

23. Mauclair. J, "Mesures de confiance en traitement automatique de la parole et applications", Université du Maine, le 5 décembre 2006.

24. Projet PAROLE, "Analyse, Perception et Reconnaissance de la parole", Lorraine, Institut National De Recherche En Informatique et en Automatique, rapport d'activités 2001.

25. Belaid. A et Belaid. Y, "Reconnaissance des formes méthodes et applications". Paris, Inter Editions 1992.

26. FABRE. X, "Exercices de reconnaissance des formes par ordinateur", Edition Masson, 1989.

27. NEY. H, Welling. H, Ortmanns. L, Beulen. S, Wessel. K et Lehrstuhl. F, "The Speech Recognition System And Spoken Document Retrieval", Industrial Electronics Society, IECON, Proceedings of the 24th Annual Conference of the IEEE, Volume 4, Issue, pp. 2022 - 2027 vol.4, Aachen, Germany,31 Aug-4 Sep 1998.

28. Bac. L.V, "Reconnaissance automatique de digits en Anglais en conditions bruitées", Université Joseph Fourier, U .F .R Informatique & Mathématiques Appliquées, Grenoble, France, 20 juin 2002.

29. Tebbi. H, "Transcription Orthographique Phonétique en vue de la synthèse de la parole à partir du texte de l'arabe standard", Mémoire de magister, Spécialité : Ingénieries des Systèmes et des Connaissances, Université de Blida, Algérie, Juin 2006.

30. Véronis. J, "Informatique et linguistique 1", Université de Provence, Centre Informatique pour les Lettres et Sciences Humaines, 1999-2001.

31. Hueber. T, "Synthèse de la parole à partir d'imagerie ultrasonore et optique de l'appareil vocal", mémoire de magistère - INSA Lyon, Ecole Centrale de Lyon, UCB Lyon, France, 2006.

32. Ozerov. A, "Représentations Robustes pour la Reconnaissance Automatique de la Parole", Stage de DESS CSA, Mars- Septembre 2003.

33. Rabiner L R et Schafer. R.W., "Digital Processing of Speech Signals", Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1978.

34. http://afcp-parole.org

35. http://perso.aricia.fr/alluin/parole/

36. Omar. M, Hamadouche. M et Loukam. M, "Conception et réalisation d'un système d'aide à la modélisation à base de modèles de Markov cachés (HMM)", Mémoire d'Ingénieur en Informatique,Université de Chlef, Algérie, juin 2004.

37. Omar. M, Hamadouche. M et Loukam. M, "EAMOMAC : Environnement d'aide à la modélisation à base de modèles de Markov Cachés", LANIA'2007, Séminaire national sur le langage naturel et l'intelligence artificielle, Chlef, Algérie, 20-21 novembre 2007.

38. Deketelaere. S, "Reconnaissance automatique de la parole", MULTITEL – Département Reconnaissance automatique de la parole, Parc Initialis-Avenue Copernic, Mons Belgique, http://www.multitel.be/ASR/download/folders/folder_fr.pdf

39. Jarifi. S, "Segmentation automatique de corpus de parole continue dédiés à la synthèse vocale", l'Ecole Nationale Supérieure des Télécommunications de Bretagne, France, 10 janvier 2007.

40. Bredin. H, "Vérification biométrique d'identité basée sur les visages parlants. Apport de la mesure de synchronie audiovisuelle face aux tentatives d'imposture élaborées", 20 juin 2007.

41. Guillaume. D, "Transformation de l'identité d'une voix", Rapport de stage du DEA ATIAM, Université de Paris VI - Université Aix Marseille II - Ecole Nationale Supérieure des Télécommunications - Institut National Polytechnique de Grenoble, 23 février 2006.

42. Doan. T T., "Verification de signature en-ligne", mémoire de fin d'études master d'informatique, 10 janvier 2006.

43. www.praat.org