الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière Électronique
Spécialité Réseau et Télécommunication

présenté par

HALIDOU Gambo Zeinabou

Segmentation automatique de la parole en phonèmes

Proposé par : Mr Abed Ahcène

Année Universitaire 2018-2019

Remerciements

En préambule à ce mémoire, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Je tiens à remercier sincèrement Monsieur ***Abed Ahcène*** *d'abord en tant qu'encadreur de ce mémoire ensuite pour ses remarques, ses conseils, sa gentillesse et la confiance dont il a fait preuve vis-à-vis de mon travail. Il m'a donné l'occasion de travailler avec autonomie, ce dont je lui suis reconnaissant.*

Je tiens à remercie aussi Monsieur ***Djendi*** *et Monsieur* ***Mehdi****, pour l'aide qu'ils m'ont apporté tout au long de mon parcours scolaire.*

Je n'oublie pas mes parents pour leur contribution, leur soutien et leur patience.

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenue et encouragé au cours de la réalisation de ce mémoire.

Merci à tous.

DEDICACE

A mes parents et à mes frères et sœurs ;

Vous vous êtes dépensés pour moi sans compter. En reconnaissances de tous les sacrifices consentis par tous et par chacun pour me permettre d'atteindre cette étape de ma vie.

A mes oncles, tantes, cousins et cousines affectueuses reconnaissances.

ملخص:

في هذا العمل، نقترح نظام آلي لتجزئة الكلام إلى وحدات صوتية للغة الإنجليزية. يعتمد هذا التقسيم على نماذج ماركوف المخفية (HMM) . الإشارات الصوتية المستخدمة في هذا العمل مستخرجة من قاعدة البيانات الأمريكية DARPA-TIMIT. لتقييم هذا النظام، أجرينا تجزئة الإشارات الصوتية تحت ظروف مختلفة. بدلالة عدد فئات النظام، عدد الحالات وكذا نوعية المعاملات. أعطى النظام النهائي قدرات بمعدل تجزئة صحيح قدره 73.95 ٪ ، تم الحصول عليه بواسطة HMM بـ 7 حالات، 2 مكونات غوسية و 14 معاملا (MFCC) .

كلمات المفاتيح: تجزئة الكلام، HMM، MFCC.

Résumé : Dans ce travail, nous proposons une segmentation automatique de la parole en phonème pour l'anglais. Cette segmentation se base sur la technique suivante: les modèles de Markov cachés HMM (Hidden Markov Models), pour modeliser les unités acoustico-phonetiques. Nous les exploitons pour classifier les signaux de parole, extraits de la base de données américaine DARPA-TIMIT, utilisée pour l'apprentissage et le test du système. Etant donnée une transcription phonétique et orthographique, notre système fournit la segmentation phonétique correspondante. Pour évaluer notre système, des expériences de segmentation et de transcription phonétique ont été effectuées dans différentes conditions. La taille de ce corpus d'apprentissage joue un role important: les performances du système ont été évaluées en fonction de ce paramètre. Le système final améliore un taux de segmentation correcte de 73.95%, obtenu par 7-HMM de 2 composantes gaussiennes pour MFCC=14.

Mots clés : Segmentation de parole ; HMM ; MFCC.

Abstract : In this work, we propose an automatic phoneme speech segmentation for English. This segmentation is based on the following technique: the Hidden Markov Models (HMM), for mode1iser acoustico-phonetic units. We use them to classify speech signals, extracted from the American DARPA-TIMIT database, used for learning and testing the system. Given a phonetic and orthographic transcription, our system provides the corresponding phonetic segmentation. To evaluate our system, segmentation and phonetic transcription experiments were performed under different conditions. The size of this learning corpus plays an important role: system performance has been evaluated as a function of this parameter. The final system improves a correct segmentation rate of 73.95%, obtained by 7-HMM of 2 Gaussian components for MFCC = 14.

Keywords : Speech segmentation ; HMM ; MFCC.

Listes des acronymes et abréviations

API	Alphabet Phonétique International
CELP	Code Linear Prediction
CG	Composante Gaussienne
DAP	Decodage Acoustico-Phonetique
dB	décibel
DCT	Discret Cosinus Transform
EM	Expectation-Maximisation
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
HTK	Hidden Markov Model ToolKit
LPC	Linear Predictive Coding
MFCC	Mel Frequency Cepstral Coefficients
MIC	Modulation à Impulsion codée
RAL	Reconnaissance Automatique du Locuteur
RELP	Residual Excited Linear Prediction
SAPh	Segmentation Automatique des Signaux en Phonèmes
TSC	Taux de Segments Corrects
VOT	Temps d'Appel Vocal
ZCR	Zero Crossing Rate
TIMIT	Texas Instruments & Massachusetts Institute of Technology

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Pourquoi la parole ? Moyen de communication naturel chez l'homme

Dans le traitement de la parole humaine, la segmentation désigne l'opération qui permet de percevoir une succession d'unités (phonèmes, syllabes, mots) dans un flux sonore continu et qui correspond au découpage du signal acoustique en unités discrètes linguistiques. (1)

La parole se distingue des autres sons par des caractéristiques acoustiques, qui ont leur origine dans le mécanisme de production. La principale fonction des sons dans une langue est d'établir des distinctions entre des unités de signification.

Les phonèmes sont les éléments sonores les plus brefs qui permettent de distinguer différents mots. Par exemple en français, les sons [p] et [b] représentent deux phonèmes différents parce que, pour un locuteur dont la langue maternelle est le français, la permutation de ces deux sons permet de distinguer de nombreuses paires de mots : pis/bis, baie/paie. (2)

La segmentation phonétique du signal de parole peut être effectuée soit manuellement par un expert humain, soit automatiquement par une méthode programmée. La conception d'un système de synthèse ou de reconnaissance automatique de parole implique une étape de segmentation des signaux de parole en phonèmes. Celle-ci est obtenue à l'aide d'une segmentation souvent manuelle et prenante. (3)

D'un point de vue qualitatif, l'examen de l'état de l'art de la segmentation de la parole donne la préférence à la segmentation manuelle. En effet, bien qu'il soit difficile d'évaluer la qualité d'une segmentation phonétique, il existe un large consensus sur le fait qu'une segmentation manuelle est plus précise qu'une segmentation automatique.

Lorsqu'il s'agit de segmenter un corpus d'une dizaine ou d'une centaine de phrases, ce processus n'est pas envisageable. Mais les besoins permanents en corpus de parole segmentée et la taille grandissante de ces corpus éliminent d'office ce type de segmentation pour son coût exorbitant. D'autant plus que des logiciels disponibles, dotés d'interfaces graphiques conviviales et interactives, représentant le signal de parole et ses caractéristiques temporelles et fréquentielles, avec une sortie audio pour l'écoute, permettent à l'expert de générer, d'une manière relativement aisée, la séquence phonétique alignée sur le signal de parole. De tous ces points découle l'intérêt majeur de la segmentation automatique de la parole. (3)

La segmentation automatique faite avec les méthodes de Markov Caché offre une précision presque égale à celle de la segmentation manuelle mais à condition de disposer et de connaitre précisément le contenu d'un corpus phonétique de parole à segmenter.

Dans le cadre de notre travail, nous proposons une Segmentation Automatique des signaux de parole en Phonèmes (SAPh), pour minimiser le temps alloué à la segmentation. Nous allons nous baser sur les méthodes de segmentation à base de modèles de Markov cachés(HMM) permettant de segmenter un énoncé de parole en phones, et ce en utilisant une transcription phonétique automatiquement produite par un système de phonétisation à partir de la représentation graphique d'un énoncé[2]. Ainsi, le chapitre 1 sera consacré au traitement du signal de la parole d'une manière générale, ensuite dans le chapitre 2 nous allons présenter les différentes étapes de la segmentation de la parole en phonème et enfin la partie expérimentale (test et implémentation) dans le chapitre 3 suivi d'une conclusion générale.

.

Chapitre 1 Traitement automatique de la parole

1.1 Introduction

Le traitement de parole est une technologie dont l'usage connaît un essor important, pour répondre notamment aux besoins des services de télécommunication, tels que les services téléphoniques ou les services de messagerie électronique. Dans le traitement de la parole humaine, la segmentation désigne l'opération qui permet de percevoir une succession d'unités (phonèmes, syllabes, mots) dans un flux sonore continu et correspond au découpage du signal acoustique en unités discrètes linguistiques.

Dans ce chapitre nous allons présenter quelques généralités sur le traitement de la parole qui peut regrouper les taches relatives à la reconnaissance de parole ; reconnaissance du locuteur et la synthèse de la parole.

1.2 Le signal de parole

La parole humaine est un flux continu constitué d'une suite de mots, eux-mêmes étant constitués d'un enchainement de phonèmes et de bruits articulatoires. La parole s'oppose à la langue par son caractère concret, individuel et créatif. La parole est en effet la réalisation phonétique de la langue résultant d'un acte psychophysiologique et volontaire de la part d'un individu.

La parole est un signal réel, continu, d'énergie finie et non stationnaire. Sa structure est complexe et variable avec le temps : tantôt périodique (plus exactement pseudopériodique) pour les sons voisés, tantôt aléatoire pour les sons fricatifs, tantôt

impulsionnelle dans les phases explosives des sons occlusifs. Cette structure reflète l'organisation temporelle des gestes de production. (3)

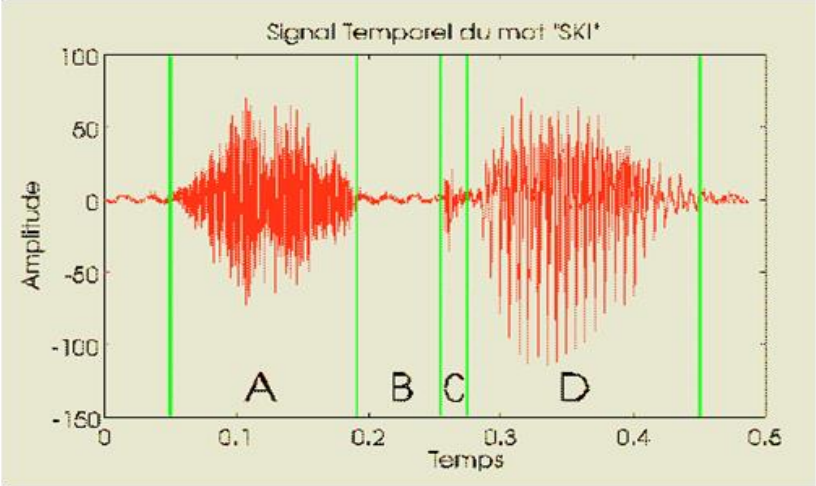


Figure 1.1 : Signal vocal. A : aléatoire, B : bruit, C : impulsion, D : pseudopériodique

La parole est très variable puisqu'un même phonème possède de nombreux paramètres qui sont fonction du locuteur.

- Intensité et Hauteur de la voix
- Type de son émis par le locuteur (chuchotement, chant, parole)
- Débit du locuteur
- Déformation du son dû à l'accent du locuteur
- Propriétés physio-acoustique de l'appareil phonatoire du locuteur
- Emotion dans la voix du locuteur (serein, pleurant, gémissant, euphorique)

Tous ces paramètres rendent le signal vocal très variable. Il est difficile d'identifier facilement les sons élémentaires. De plus, lorsqu'il y a plus d'un locuteur, ce travail devient très sérieux avec le mélange de signaux. Un son se définit classiquement au moyen de son amplitude, de sa durée, et de son timbre. Le traitement du signal a pour but précisément de quantifier ces trois grandeurs pour faire correspondre à l'onde sonore (temporelle) une description multidimensionnelle. (4)

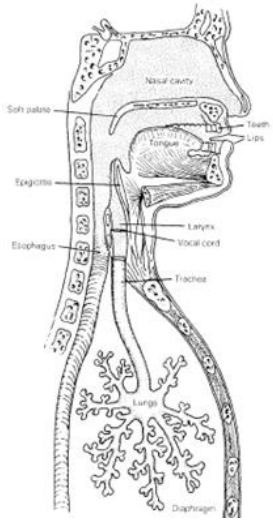
1.2.1 Phonème et Phonétique

La phonétique est le domaine de la linguistique qui a pour objet l'étude des langues naturelles dans leurs dimensions sonores. Le phonème est la plus petite unité discrète que l'on puisse isoler par segmentation dans la chaîne parlée. Un phonème est en réalité une entité abstraite, qui peut correspondre à plusieurs sons. Il est en effet susceptible d'être prononcé de façon différente selon les locuteurs ou selon sa position et son environnement au sein du mot. Les phones sont d'ailleurs les différentes réalisations d'un phonème. (5)

1.2.2 Production de la parole

La parole est un phénomène acoustique qui se distingue des autres sons par des caractéristiques liées aux mécanismes de sa production par l'appareil phonatoire. Ce dernier fait intervenir divers éléments : l'air, comme source d'énergie ; les cordes vocales, comme principal organe vibratoire ; la langue et les lèvres, comme organes vibratoires accessoires ; les cavités buccale et nasale, comme caisses de résonance. (3)

Habituellement, la parole est produite grâce à une pression appliquée sur les poumons puis par la modulation du courant d'air qui entre dans le canal vocal, ensemble de mouvements coordonnés qui permettent la phonation. Toutefois, la production de la parole est possible sans utiliser les poumons et la glotte ; le discours **laryngé** utilise les parties hautes du canal vocal.



L'appareil vocal humain:

source excitatrice	poumons, trachée
élément vibrant	cordes vocales
résonateurs	cavités buccales et nasales
articulateurs	dents, lèvres, langue

Figure 1.2 : Appareil vocal humain (6)

La parole est le résultat de l'excitation des cavités **supra glottiques** (nasales ou orales) par une ou deux sources acoustiques.la première, essentielle, génère une onde de débit, c'est la source laryngienne ; on peut la considérer comme quasi périodique. L'autre peut s'ajouter ou se substituer à la première : il s'agit cette fois de bruits d'explosion ou de friction qui peuvent naitre à l'intérieur du conduit vocal (de la glotte aux lèvres), s'il y a un rétrécissement nettement marqué ou si **une occlusion** se relâche brusquement. Cette (ces) source(s) va (vont) exister le conduit vocal, dont la disposition dépend de l'articulation. (7)

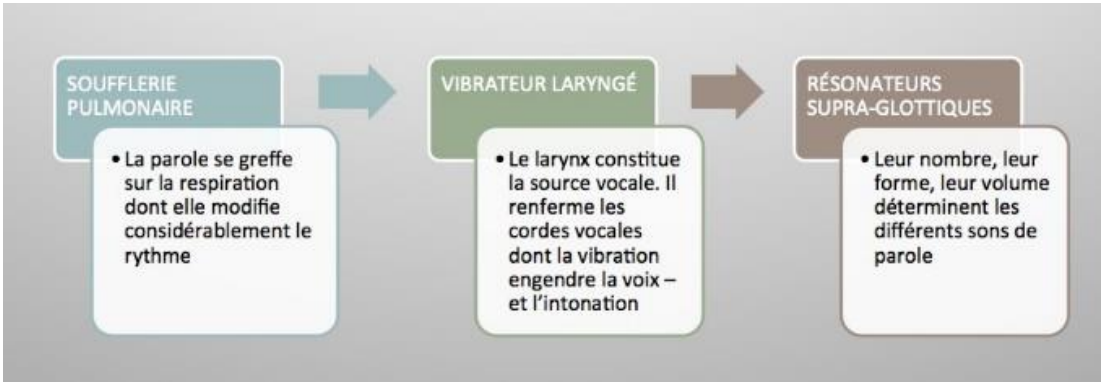


Figure 1.3 : Organes intervenant dans la production de la parole

1.2.3 Perception de la parole (5)

La **perception de la parole** est le processus par lequel les humains sont capables d'interpréter et de comprendre les sons utilisés dans le langage. Une vibration

mécanique de la matière et de l'air qui met en alternance le tympan ou le micro ne constitue pas en elle-même un son. Car c'est dans le cerveau que naît et se forme le son. Le son n'existe pas en-dehors de notre cerveau, de nous-même. L'oreille recueille les vibrations de l'air, les transforme en impulsion électrique au moyen des cellules nerveuses, impulsion qui est perçue et interprétée en son par le cerveau.

L'oreille est le principal organe responsable du processus de perception. Le fonctionnement de l'oreille a deux parties: le comportement de l'appareil mécanique et le traitement neurologique des informations acquises. Entendre, l'un des cinq sens de l'humain, c'est la capacité de percevoir le son en détectant les vibrations via l'oreille d'un organe.

Le processus mécanique de l'oreille est le début du processus de perception de la parole. Les processus complexes par lesquels un auditeur « comprend » un message oral émis par un locuteur peuvent être fonctionnellement décomposés en deux grandes phases : dans une première phase, l'oreille transforme l'information contenue dans le signal acoustique et la transmet au cerveau par l'intermédiaire du nerf auditif (sélection, traitement), la deuxième phase correspond à la reconstitution du message linguistique sur la base de cette représentation du signal de parole. Cette deuxième phase peut elle-même être décomposée en deux niveaux, l'un correspondant à l' « interprétation » d'indices fournis à l'issue du prétraitement auditif sans référence à la signification, et l'autre réalisant l'accès au sens.

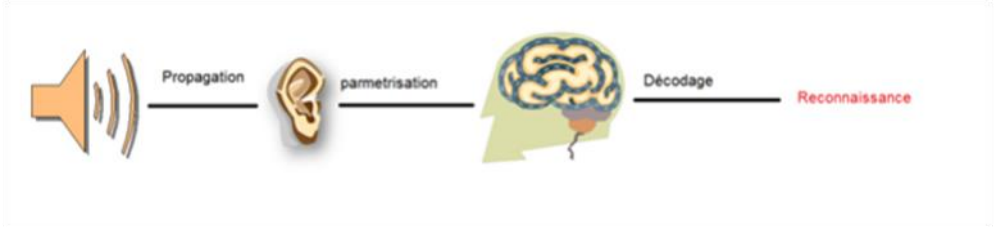


Figure 1.4 : Perception et analyse du son par l'être humain

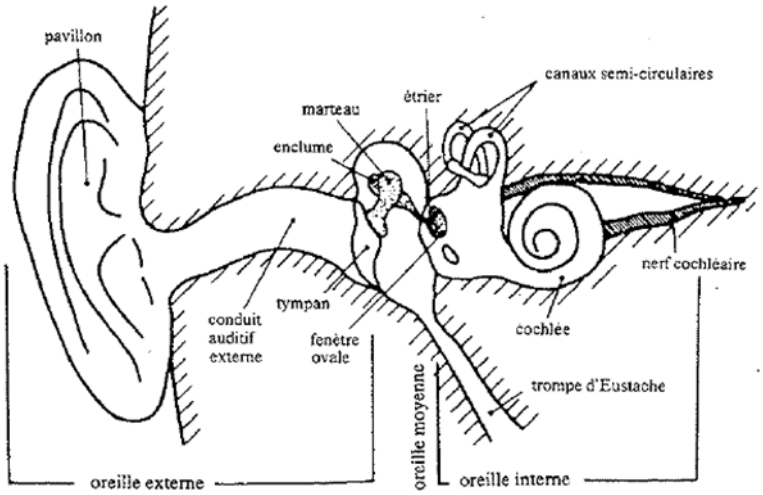


Figure 1.5 : Schéma de l'anatomie du système auditif périphérique (2)

1.3 Analyse des signaux de parole

Les techniques d'analyse du signal de parole décrites dans ce chapitre sont désormais éprouvées et offrent une base solide aux techniques de traitement. Ainsi de nombreux efforts ont été fournis en recherche de traitement du signal dans le cadre de la reconnaissance du locuteur, en synthèse ou en reconnaissance de la parole. Les mécanismes de production ainsi que les paramètres caractéristiques du signal présentés ici proviennent de ces efforts.

1.3.1 Numérisation de la Parole

La parole est produite par l'articulation des membres phonatoires de l'homme et prend une forme analogique apériodique ; ce qui est impossible pour que la machine puisse l'interpréter ou le prédire car elle ne comprend que du numérique. Pour cela on doit faire un traitement de numérisation sur ce signal. L'une des méthodes les plus utilisés dans la numérisation est la méthode Delta ou MIC qui consiste en trois étapes : l'échantillonnage, la quantification et le codage. (5)

1.3.2 Energie

L'énergie du signal est un indice qui peut par exemple contribuer à la détection du voisement d'un segment de parole. L'énergie totale E_0 est calculée directement dans le domaine temporel sur une trame de signal $\boldsymbol{s(n)}$ avec $\boldsymbol{n}$ entre 0 et $\boldsymbol{N-1}$ comme :

$$E_0 = \frac{1}{N}\sum_{n=0}^{N-1}|S(n)|^2 \quad (1)$$

L'énergie ainsi obtenue est sensible au niveau d'enregistrement; on choisit en général de la normaliser, et d'exprimer sa valeur en décibels par rapport à un niveau de référence. L'énergie est très importante dans la reconnaissance des émotions, elle est incluse dans tous les outils d'extraction de paramètres (8) .

1.3.3 Taux de passage par zéro (9)

Le ZCR (Zéro Crossing Rate) est le taux de passage par zéro. Pour un signal échantillonné, il y a passage par zéro lorsque deux échantillons successifs sont de signes opposés ; le taux de passage par zéro court terme peut être estimé par la formule :

$$Z(n) = \sum_{m=-N/2}^{N/2} |\, sgn\, x(n+m) - sgn\, x(n+m-1)|\ .w(m) \quad (2)$$

Où

$$w(n) = f(x) = \begin{cases} 1/N, & -N/2 \le n \le N/2 \\ 0, & ailleurs \end{cases} \quad (3)$$

Les brusques variations du ZCR sont significatives de l'alternance voisée / non voisée donc de présence de parole. Il est clair que les valeurs de Z(n) sont normalement plus élevées pour les sons non voisés que pour les sons voisés (Le voisement est une qualité (ou propriété) de certains sons de la parole. Un son est voisé si sa production s'accompagne d'une vibration des cordes vocales et sinon, il est non voisé. On utilise aussi couramment les termes de sonore et sourd pour désigner cette opposition).

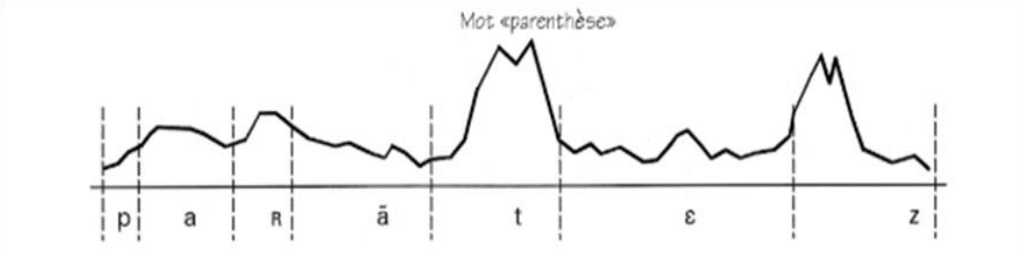


Figure 1.6 : Taux de passage par zéro pour le mot « parenthèse »

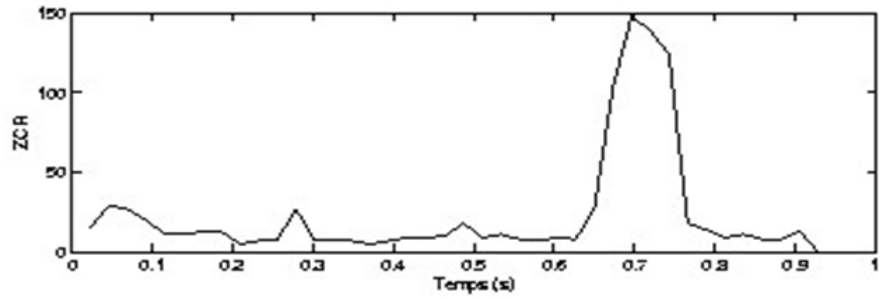


Figure 1.7 : ZCR d'un signal de parole de 1s (les pics correspondent aux zones non voisées).

Il est à remarquer que le calcul du taux de passage à zéro est extrêmes sensible à la présence d'une composante continue dans le signal de parole. Une estimation correcte de cette grandeur ne peut être faite que sur un signal à moyenne nulle.

Remarque : L'énergie est assez liée au ZCR. En effet, l'énergie varie beaucoup pour la parole alors qu'elle est plutôt stable pour la musique. Elle permet donc, comme le ZCR, de discriminer la parole et la musique mais aussi de détecter les silences. Souvent l'énergie est couplée au ZCR : un taux de passage par zéro faible et une énergie forte étant synonyme d'un son voisé alors qu'un taux de passage par zéro élevé induit une zone non voisée.

1.3.4 Fréquence fondamentale et formants

La fréquence fondamentale est un composant de basse fréquence de la parole, résultant de la vibration des cordes vocales, permettant la perception de la hauteur tonale de la voix d'un individu. 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. (10)

La fréquence fondamentale F0 (ou pitch [période des sons voisés]) joue un rôle important dans la parole. C'est elle qui véhicule une grande partie de l'information prosodique. L'intensité de la voix et les durées successives des syllabes complètent ces informations. D'une manière générale, la prosodie, qui peut être considérée comme l'effet des différentes variations de la fréquence fondamentale F_0, de l'intensité et de la durée, peut faire ressortir bien des caractéristiques du locuteur, comme son genre, ses origines géographiques et culturelles, ses émotions, etc. mais participe aussi à la

caractérisation de la langue elle-même, par la manière dont elle est utilisée pour différencier les divers éléments syntaxiques comme les énoncés (interrogatifs, exclamatifs ou déclaratifs), l'importance de certains mots, ou bien même pour caractériser les différences lexicales entre les mots.

On appelle formant les maxima locaux (pics) de la fonction de transfert notés F1, F2, F3… et correspondent aux zones de renforcement maximal. Les cavités supra-glottiques ont la capacité de neutraliser certaines harmoniques et d'en mettre d'autres en évidence par un simple changement de configuration. Lorsque l'on prononce, sur une note constante ou à une hauteur de voix constante, des voyelles aussi différentes que " oe i u ", c'est le procédé d'atténuation et de renforcement qui entre en jeu et qui est responsable de l'apparition du timbre propre à chacune des voyelles. Donc le conduit vocal possède une fonction de transfert (filtre) appliquée à une source produit le phonème. (10)

1.3.5 MFCC

L'objectif de cette phase de reconnaissance est d'extraire des coefficients représentatifs du signale de la parole. Ces coefficients sont calculés à intervalles réguliers. En simplifiant les choses, le signal de la parole est transformé en une série de vecteurs de coefficients, ces coefficients doivent représenter au mieux ce qu'ils sont censé modéliser et doivent extraire le maximum d'informations utiles pour la reconnaissance. Parmi les coefficients les plus utilisés et qui représentent au mieux le signal de la parole, nous trouvons les coefficients ceptraux, appelés également ceptres.

Les coefficients MFCC sont un type de coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole. Le codage MFCC utilise une échelle fréquentielle non-linéaire. (8)

1.3.6 LPC (11)

Le codage prédictif linéaire (LPC, Linear Predictive Coding) est une méthode de codage et de représentation de la parole. Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus linéaire. Il s'agit donc de prédire le signal à un instant n à partir des p échantillons précédents. La parole n'étant

cependant pas un processus parfaitement linéaire, la moyenne que constitue la somme pondérée du signal sur pas de temps introduit une erreur qu'il est nécessaire de corriger par l'introduction du terme $e(n)$. Le codage par prédiction linéaire consiste donc à déterminer les coefficients a_k qui minimisent l'erreur $e(n)$, ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage.

La méthode du codage par prédiction linéaire est tout autant utilisée en RAP qu'en compression pour le transfert de la voix par téléphone ou radio. Elle n'est cependant pas parfaite puisque l'erreur de prédiction peut être importante sans qu'il soit possible, par cette méthode, de la corriger. La méthode RELP, Residual Excited Linear Prediction, permet de réduire une partie de cette erreur. Le principe consiste à comparer, lors de la prédiction linéaire, le signal obtenu avec le signal original. L'erreur, obtenue par soustraction, représente la partie du signal original que le prédicteur n'arrive pas à modéliser. Dans la méthode RELP, l'erreur résiduelle est passée dans un filtre passe-bas permettant de conserver l'erreur effectuée dans la seule bande fréquentielle allant de 0 à 1000 hertz. La sortie du filtre est alors codée et passée au receveur qui peut alors reconstruire un signal à partir de la prédiction et de l'erreur observée. Pour pallier le problème de l'erreur résiduelle, d'autres méthodes fondées sur la prédiction linéaire ont été développées. Ainsi la méthode CELP, Code Linear Prediction, permet d'effectuer une compression de la parole par codage d'une trame vis-à-vis de références stockées dans un corpus. Ainsi, une trame de parole sera codée selon une combinaison linéaire de certaines trames du corpus et c'est cette combinaison linéaire qui sera considérée à la place de la trame dans les traitements ultérieurs. Cette méthode de codage de la parole est surtout employée pour la compression et la transmission de la parole à de faibles débits.

1.4 Traitement automatique de la parole

Le traitement de la parole est une discipline technologique dont l'objectif est la captation, la transmission, l'identification et la synthèse de la parole. A cet égard, l'analyse et la synthèse sont duales : l'une opère sur le signal vocal et fournit une description dans un espace de représentation, généralement non temporel et la seconde part de cette description pour produire le signal.

Mais on peut aussi traiter le signal pour réduire la redondance, dans le but soit de trouver des paramètres pertinents pour la reconnaissance, soit de comprimer l'onde sonore avant stockage ou transmission.

C'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 interviennent à des degrés divers dans les solutions apportées .

1.4.1 Reconnaissance de la parole (12)

La Reconnaissance de la parole continue consiste à transformer le "flot" acoustique du signal de parole en une représentation symbolique. Cette représentation doit être caractéristique du contenu linguistique. La reconnaissance de la parole continue peut être basée directement sur une comparaison de formes nouvelles avec des références des mots à reconnaître (ex., description complète des mots en termes de modèles acoustiques), ou bien sur l'identification d'un ensemble d'unités élémentaires (ex., phonèmes, diphones, syllabes). Dans le premier cas, il s'agit d'une reconnaissance globale, dans le second cas, d'une reconnaissance analytique.

Objectif: retrouver une information à partir de la voix (inverse de la synthèse vocale dont l'Entrée représente des données provenant d'un analyseur ,signal de voix déjà traité et la Sortie: texte).Cela dit les connaissances nécessaires pour la compréhension de la parole sont énormes, ainsi nous allons nous contenter d'étudier l'aspect reconnaissance de la parole qui a pour objectif de décoder le signal de la parole en unités de bases (phonèmes, mots …) sans en donner une signification (sans comprendre le sens des phrases construites). De nos jours, les systèmes de reconnaissance de la parole ont évolué et utilisent non seulement des connaissances en linguistique (Phonétiques, Phonologiques, Prosodiques, Lexicales…) mais aussi des connaissances dans les domaines : Traitement du signal, Reconnaissance des formes

Parole: Flux continu constitué d'une suite de mots, eux-mêmes constitués d'un enchainement de phonèmes et de bruits articulatoires Phonèmes: plus petite unité discrète ou distinctive (c'est-à-dire permettant de distinguer des mots les uns des autres) que l'on puisse isoler par segmentation dans la chaîne parlée.

a La méthode globale

Cette méthode considère le plus souvent le mot comme unité de reconnaissance minimale, c'est-à-dire indécomposable. Dans ce type de méthode, on compare globalement le message d'entrée (mot, phrase) aux différentes références stockées dans un dictionnaire en utilisant des algorithmes de programmation dynamique. Cette méthode a pour avantage d'éviter l'explicitation des connaissances relatives aux transitions qui apparaissent entre les phonèmes. Ce type de méthode est utilisé dans les systèmes de reconnaissance de mots isolés, reconnaissance de parole dictée avec pauses entre les mots… et présente l'inconvénient de limiter la taille du dictionnaire.

b La méthode analytique

Cette méthode fait intervenir un modèle phonétique du langage. Il y a plusieurs unités minimales pour la reconnaissance qui peuvent être choisies (syllabe, demi-syllabe, diphone, phonème, phone homogène, etc.). Le choix parmi ces unités dépend des performances des méthodes de segmentation utilisées. La reconnaissance dans cette méthode, passe par la segmentation du signal de la parole en unités de décision puis par l'identification de ces unités en utilisant des méthodes de reconnaissance des formes (classification statistique, réseau de neurones, etc.) ou des méthodes d'intelligence artificielle (systèmes experts par exemple). Cette méthode est beaucoup mieux adaptée pour les systèmes à grand vocabulaire et pour la parole continue. Les problèmes qui peuvent apparaître dans ce type de système sont dus en particulier aux erreurs de segmentation et d'étiquetage phonétique. C'est pourquoi le DAP (Décodage Acoustico-Phonétique) est fondamental dans une telle approche.

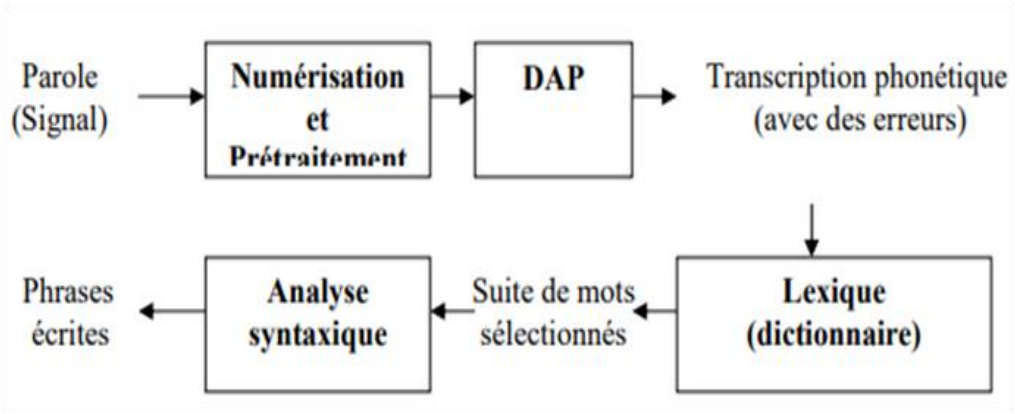


Figure 1.8 : Schéma général d'un système de reconnaissance de parole continu

Un système de reconnaissance de la parole peut être utilisé par une seule personne (mono locuteur), plusieurs personnes (multi locuteurs) ou tout le monde (indépendant du locuteur).

1.4.2 Reconnaissance de locuteur

Il convient dans ce domaine de recherche de reconnaitre non pas ce qui a été dit, mais l'identité de la personne qui parle, à partir de son **empreinte** vocale.

La Reconnaissance Automatique du Locuteur (RAL) vise à déterminer, automatiquement, si un échantillon de voix a été prononcé par une personne donnée. Les tâches courantes en RAL peuvent être classées en deux grandes catégories : Les tâches relevant de l'identification du locuteur consistent à rechercher, parmi un ensemble de locuteurs connus, le locuteur possédant la référence la plus proche d'un message vocal donné. L'identification du locuteur peut être réalisée en ensemble fermé (le message vocal à identifier a été prononcé par un des locuteurs d'un ensemble, fermé et sont tous connus du système) ou en ensemble ouvert (un locuteur inconnu a pu prononcer le message). La vérification du locuteur revient à évaluer l'hypothèse qu'un locuteur donné ait prononcé le message vocal considéré. (2)

Une autre caractéristique importante concerne le mode de participation des locuteurs. Dans la majeure partie des applications et des travaux de recherche, les

locuteurs collaborent avec le système automatique : ils prononcent ce que demande le système et/ou ils cherchent à être reconnus par celui-ci.

La nature des messages vocaux manipulés par le système peut également être plus ou moins contrainte : le système est dépendant du texte si les locuteurs sont tenus de prononcer un texte précis, indépendant du texte dans le cas contraire.

Un système de reconnaissance du locuteur comporte quatre éléments principaux (12):

- Un module d'acquisition et de paramétrage du signal. Celui-ci a en charge de produire une suite de vecteurs de paramètres, dits paramètres acoustiques, représentant le message vocal sous une forme exploitable par le système. Les méthodes employées dérivent des analyses de Fourier (FFT, analyse Cepstrale, PLP...)
- Un modèle d'apprentissage. Celui-ci crée une référence vocale pour un locuteur à partir d'un échantillon de sa voix. L'approche statistique (GMM) est majoritairement utilisée.
- Un module de calcul de ressemblance. Ce module est utilisé durant la reconnaissance. Il a en charge de calculer la ressemblance entre un échantillon de signal et une référence correspondant à une personne donnée.
- Enfin, un module de décision utilise durant la phase de reconnaissance les sorties du module précédent pour prendre une décision.

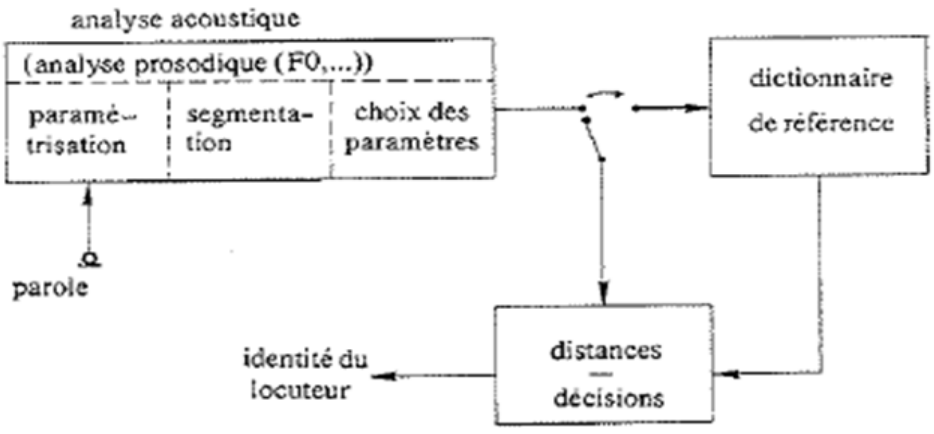


Figure 1.9 : Schéma d'un système de reconnaissance du locuteur (2)

Un étage d'analyse acoustique comprenant un module de paramétrage utilisant des méthodes d'analyse classique ou adapté à cette tâche, auquel on pourra ajouter une analyse du fondamental, et de la courbe d'énergie, suivi éventuellement d'un module de segmentation et d'un module de choix des paramètres pertinents. Lors de l'apprentissage, ces paramètres sont rangés en mémoire avec le nom du locuteur auxquels ils correspondent.

A la reconnaissance, un module de comparaison va calculer la distance entre les paramètres du signal vocal prononcé, et ceux des locuteurs présents dans la mémoire (ou du locuteur dont l'identité a été donnée, dans le cas d'un système de vérification).

Des seuils de rejet permettent de prendre une décision sur l'identité du locuteur.il faut noter que ces seuils de rejet doivent tenir compte des variations de prononciation d'un même locuteur (variation interlocuteur). Cette opération de comparaison doit, dans l'idéal, être capable de trouver plus voisines deux prononciations très différentes (voix criée et chuchotée), de mots différents (si le système est indépendant du texte) par un même locuteur, que deux prononciations normales d'un même mot par deux locuteurs différents. (2)

Remarque : La nature très variable du signal de parole est un facteur délicat à gérer. La voix évolue avec l'âge, l'état physiologique ou pathologique du locuteur. Le canal de transmission de l'information joue un rôle important en RAL : en diminuant la qualité du matériel de comparaison (bande passante, bruits...) et en influençant la décision (le système ne doit pas prendre en compte les caractéristiques du canal pour reconnaître une personne mais les spécificités du locuteur). Enfin, et ce n'est pas forcément la moindre des difficultés, l'évaluation des performances n'est pas aisée. Cette évaluation nécessite des bases de données spécifiques, de grande taille et contenant un sous ensemble représentatif des difficultés rencontrées. En particulier, simuler un imposteur capable " d'imiter " une personne donnée n'est pas trivial (actuellement les tests d'imposture sont majoritairement réalisés en confrontant un

locuteur de la base de test à tous les autres locuteurs de la base, sans modifier les éléments de test). (12)

1.4.3 Synthèse de la parole

La synthèse vocale est une technique informatique de synthèse sonore qui permet de créer de la parole artificielle à partir de n'importe quel texte. Pour obtenir ce résultat, elle s'appuie à la fois sur des techniques de traitement linguistique, notamment pour transformer le texte orthographique en une version phonétique prononçable sans ambiguïté, et sur des techniques de traitement du signal pour transformer cette version phonétique en son numérisé écoutable sur un haut-parleur. (13)

Les premiers synthétiseurs de parole furent construits pour vérifier l'hypothèse suivant laquelle le riche spectre de la parole pouvait être, sans une dégradation de l'intelligibilité du message parlé, réduit à un petit nombre de composantes. La parole est très redondante aussi bien sur le plan du signal lui-même que sur le plan linguistique. Nous savons que 40 000 unités d'information par seconde environ sont transmises par une ligne téléphonique classique au cours d'une conversation alors que notre cerveau, selon certaines hypothèses, ne pourrait décoder, en fin de traitement qu'un maximum de 50 unités d'information par seconde, ces 50 unités d'information étant porteuses de données sur le message ainsi que sur la qualité et sur les caractéristiques individuelles de la voix. Le rôle et l'importance de telle ou telle composante du signal peuvent être étudiés au moyen de la synthèse. (14)

Pour que la synthèse vocale soit utilisable, il ne suffit pas que le texte lu, soit intelligible, c'est-à-dire articulé suffisamment correctement pour pouvoir être décodé par l'auditeur. Il faut aussi que la voix ait une qualité suffisante, c'est-à-dire soit suffisamment naturelle pour être acceptable par l'auditeur.

Cela implique non seulement une restitution correcte de chaque phonème, mais aussi une coarticulation naturelle de ces phonèmes et une mélodie naturelle. La lecture d'un texte numérisé par l'ordinateur nécessite trois étapes (15):

- un prétraitement du texte destiné à en éliminer les "anomalies"

- Convertir tous les signes graphiques qui n'ont pas de fonction phonographique directe en suite de graphèmes à fonction phonographique : il s'agit des abréviations graphiques diverses, des logogrammes (y compris les chiffres)

- la poétisation du texte (au sens large), c'est-à-dire l'élaboration, à partir de la représentation graphique, d'une représentation phonique incluant les faits segmentaux et les faits suprasegmentaux incluant la transformation des suites de graphèmes en suites de phonèmes et le calcul de la prosodie (pauses, courbes mélodiques, accentuation, durée des phonèmes). Ce calcul nécessite une analyse linguistique plus ou moins approfondie du texte. Cette analyse inclut les aspects suivants (qui sont partiellement liés les uns aux autres) :

  - Une analyse morphologique destinée à l'identification des mots, analyse des mots inconnus. Le programme peut s'appuyer sur un lexique ou mettre en jeu des règles d'analyse morphologique.

    Ex. des sigles : doivent-ils être épelés ou lus ?

  - Une analyse syntaxique permettant l'étiquetage des unités repérées, la désambiguïsation des homographes et surtout la délimitation des syntagmes indispensable pour le calcul de la prosodie.
    Ex. <il le boit>. <le> peut être pronom ou déterminant. l'identification de <boit> comme verbe permet l'identification de <le> comme pronom. A la différence de <le> dans <le monsieur>

  - Une analyse des topogrammes : Ex. du point, qui peut marquer la fin d'une phrase, mais qui peut être aussi le signe terminal d'une abréviation graphique à oraliser (M.), un séparateur dans des numéros de téléphone (01.03.03.03.03), il sert alors de démarcatif entre groupes de chiffres et n'a pas de réalisation phonétique propre, ou un séparateur dans les numéros de versions de logiciels (Version 10.6), il doit alors être oralisé

- la synthèse vocale consiste en la production des sons correspondant à la représentation phonétique élaborée. Elle consiste, dans son principe, en la

superposition de deux types de données : les données segmentales (concaténation de phonèmes) et les données suprasegmentales.

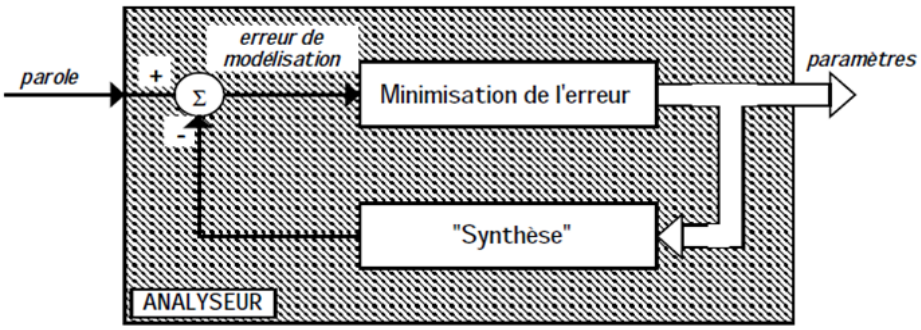


Figure 1.10 : Schéma de principe d'un analyseur de parole (2)

Parmi les applications, on peut citer ,la vocalisation d'écrans informatiques pour les personnes aveugles ou fortement malvoyantes1 (lecteur d'écran), les applications embarquées (automobile, traducteurs ,automatiques de poche) ; les télécommunications (services de consultation de courrier électronique par téléphone, serveurs vocaux interactifs, livres et journaux parlants) et du multimédia (jeux informatiques, aide aux handicapés, " machines à lire " avec scanner et OCR pouvant servir aux aveugles et malvoyants) ; ainsi que de nombreuses applications de serveurs vocaux téléphoniques, comme les annuaires vocaux de grande taille, où la synthèse vocale est la seule technique viable pour permettre la restitution sonore des noms et des adresses des abonnés (2).

1.5 Conclusion

Ce chapitre a été consacrée principalement à fournir au lecteur : les concepts théoriques de la production de la parole et les outils mathématiques nécessaires pour son analyse et son traitement numérique.

Les principaux problèmes posés en traitement du signal proviennent de la dualité « source /produit » de l'appareil vocal humain, de la grande dynamique et de la variété de la voix, ainsi que des variations rapides de la parole. De même les phonèmes dans les mots ont des images temporelles différentes : il s'agit d'une variabilité contextuelle (le locuteur étant le même) .

Il est donc de la première importance de choisir une méthode d'analyse en fonction des objectifs fixés : est-ce pour la reconnaissance ? (problème de la pertinence des paramètres extraits de leur invariance) pour le codage ? (problème de la réduction du débit, du taux de distorsion,) est ce pour la synthèse ? (problème du naturel de l'intelligibilité).

Dans le chapitre suivant nous essayons d'introduire la segmentation automatique de parole en phonèmes. Le système proposé est basé sur les modèles de Markov Cachés.

Chapitre 2 Système de segmentation de la parole

2.1 Introduction

Lorsque l'on écoute de la parole, on a l'impression qu'elle est composée d'un enchaînement de sons distincts. La séparation des mots à partir du flux de la parole est indispensable à la compréhension du discours et à l'accès lexical. La plupart des systèmes d'écriture possèdent des espaces blancs entre les mots. Dans le signal de parole, il n'existe pas d'indices clairs et univoques qui permettent de marquer le début et la fin des mots. La parole est dite « continue » : la segmentation est donc une étape majeure dans la reconnaissance des mots parlés.

La segmentation est formulée comme une estimation des emplacements et des durées de parole et des segments autres que de la parole signal vocal d'entrée. Toutes les composantes de la parole sont nichées ensemble pour former un ensemble contenant un signal ainsi que des composants analogues au bruit, tandis que les composants non-parole résidant dans la durée d'inactivité entre les composants de la parole forment un autre ensemble contenant seuls les composants ressemblant à du bruit.

La segmentation correspond au découpage du signal acoustique en unités discrètes linguistiques. Des difficultés apparaissent notamment dans plusieurs contextes comme l'acquisition de la langue maternelle, la reconnaissance de mots parlés et l'acquisition d'une seconde langue, du fait de la double articulation. Cependant, des stratégies ont été mises en évidence pour pallier à ce problème de segmentation(16).

.

Il serait alors possible de mettre au point un système artificiel capable d'identifier avec certitude la chaîne de phonèmes correspondant au signal acoustique qui lui est fourni en entrée. Supposons maintenant que ce système dispose également de procédures lui permettant de traduire la chaîne de phonèmes en représentations phonologiques. Ce système disposerait alors de représentations sous-jacentes directement appariables avec les éléments stockés dans son lexique. On peut penser qu'il suffirait alors d'envisager, à partir de cette chaîne d'entités phonologiques sous-jacentes, l'ensemble des suites de mots possibles pour retrouver la séquence de mots effectivement produite par le locuteur. De manière simplifiée, c'est l'une des stratégies qui ont été proposées dans plusieurs modèles de la segmentation lexicale en psycholinguistique. Cette procédure est aussi largement utilisée en Reconnaissance.

2.2 Spécificité de la langue Anglaise (17)

Depuis quelques années, on n'envisage que l'une des sources essentielles d'information pour la localisation des frontières de mots dans un signal de parole est l'identification même des mots qui le constituent. Certains modèles d'accès au lexique se sont alors présentés comme un choix particulièrement bien adapté au problème de la segmentation lexicale d'un signal de parole continue.

La langue anglaise se caractérise par l'existence de paramètres accentuels liés aux représentations lexicales (6). Dans les mots anglais, on observe une alternance de syllabes accentuées (strong syllables, syllabes fortes) et non accentuées (weak syllables, syllabes faibles). Cette alternance étant déterminée par les représentations lexicales, elle a un caractère distinctif : la position de l'accent dans une séquence de phonèmes peut déterminer sa signification. Cet accent lexical présente cependant une régularité importante en ce qui concerne la position qu'il occupe dans les mots. En anglais (18), 80% des mots d'un lexique informatisé portent l'accent sur leur syllabe initiale. Dans un système de segmentation de la parole en mots qui chercherait des informations acoustiques aptes à indiquer des frontières de mots, la prise en considération de cette régularité permettrait de privilégier les syllabes fortes comme des débuts de mots.

Les travaux présentés montrent que certaines régularités allophoniques peuvent être utilisées par des locuteurs de la langue comme indicateurs de frontières de mots. En anglais, il existe par exemple une tendance à prononcer les phonèmes en position initiale de mots avec des caractéristiques particulières qui les distinguent des mêmes phonèmes prononcés en position finale ou médiane. Ces différences de prononciation peuvent être utilisées par des locuteurs pour identifier une séquence de parole ambiguë (8).

Les auteurs montrent, dans une analyse acoustique de séquences lexicalement ambiguës, que le signal de parole est porteur d'indices différenciant les phonèmes selon qu'ils sont prononcés en début de mot ou dans d'autres positions. Ainsi, les consonnes occlusives sont glottalisées ou laryngealisées en position initiale : le /½/ médian de 'no notion',qui se trouve à l'initiale du second mot a des caractéristiques acoustiques différentes du /½/ médian de 'known ocean' qui est localisé à la fin du premier mot. Cette modification des caractéristiques acoustiques des consonnes en fonction de leur position lexicale permet à des auditeurs d'identifier correctement la séquence effectivement produite malgré son ambiguïté phonologique (avec 70 à 90% de taux de correspondance entre la séquence lexicale effectivement produite et la réponse des auditeurs). Par ailleurs, c'est la laryngealisation ou la glottalisation de la consonne en début de mot qui constitue un indice pertinent ; l'absence de cette caractéristique n'induit pas nécessairement que le phonème soit en position médiane ou finale de mot.

Outre les variations ***allophoniques*** dépendantes des frontières lexicales, il existe aussi des modifications acoustiques des phonèmes en fonction de leur position dans la syllabe [8].il a été proposé d'intégrer des connaissances sur les régularités des alternances allophoniques liées à la structure syllabique dans les processus de 'segmentation' (plus précisément de parsing) lexicale. Ce type d'informations semble effectivement être mis en œuvre par le système de traitement de la parole.

2.2.1 Les consonnes

En phonétique articulatoire , une consonne est un son de parole articulé avec une fermeture complète ou partielle du tractus vocal (19). Des exemples sont [p] ,

prononcés avec les lèvres; [t] , prononcé avec le devant de la langue; [k] , prononcé avec le dos de la langue; [h] , prononcé dans la gorge; [f] et [s] , prononcés en forçant de l'air à travers un canal étroit (fricatives); et [m] et [n] , dans lesquels l'air circule par le nez (les nasales).

Comme le nombre de sons possibles dans toutes les langues du monde est beaucoup plus grand que le nombre de lettres d'un alphabet , les linguistes ont mis au point des systèmes tels que l' alphabet phonétique international (IPA) pour attribuer un symbole unique et non ambigu à chaque consonne attestée. En fait, l' alphabet anglais a moins de lettres de consonnes que l'anglais a des sons de consonnes, aussi des digrammes comme "ch", "sh", "th" et "zh" sont utilisés pour étendre l'alphabet, et certaines lettres et digraphes représentent plus que une consonne. Par exemple, le son orthographié "th" dans "this" est une consonne différente de celle du "th" son dans "thin".

Les consonnes se divisent en deux groupes (selon que l'on fasse vibrer ou non les cordes vocales) : sourdes, sans beaucoup de vibrations, et sonores, avec plus de vibrations. Les voyelles, toujours sonores, prononcées avec les consonnes cachent un peu le caractère sourd des consonnes.

Chaque consonne parlée peut être distinguée par plusieurs caractéristiques phonétiques :

- Le mode d'articulation est la façon dont l'air s'échappe du conduit vocal lorsque le son de la consonne ou de l'approximant (comme une voyelle) est émis. Les manières comprennent les arrêts, les fricatives et les nasales.
- Le lieu d'articulation est l'endroit où se produit l'obstruction de la consonne dans le conduit vocal et quels organes de la parole sont impliqués. Les endroits incluent bilabial (les deux lèvres), alvéolaire (langue contre la crête des gencives) et vélaire (langue contre palais mou). De plus, il peut y avoir un rétrécissement simultané en un autre lieu d'articulation, tel que la palatalisation ou la pharyngalisation . Les consonnes ayant deux lieux d'articulation simultanés sont dites co-articulées .

- La phonation d'une consonne est la façon dont les cordes vocales vibrent pendant l'articulation. Lorsque les cordes vocales vibrent complètement, la consonne s'appelle une voix ; quand ils ne vibrent pas du tout, ce n'est sans voix .
- Le temps d' appel vocal (VOT) indique le moment de la phonation. L'aspiration est une caractéristique de VOT.
- Le mécanisme de flux d'air est la façon dont l'air qui circule dans le conduit vocal est alimenté. La plupart des langues ont des consonnes exclusivement pulmonaires, qui utilisent les poumons et le diaphragme, mais les éjectifs , les clics et les implosifs utilisent des mécanismes différents.
- La longueur est la durée de l'obstruction d'une consonne. Cette caractéristique est très différente en anglais, comme dans "tout à fait" [hoʊlli] contre "sainte" [hoʊli], mais les cas se limitent aux limites d'un morphème.
- La force articulatoire est la quantité d'énergie musculaire impliquée. Cela a été proposé à maintes reprises, mais aucune distinction reposant exclusivement sur la force n'a jamais été démontrée.

Toutes les consonnes anglaises peuvent être classées par une combinaison de ces caractéristiques, comme "arrêt alvéolaire sans voix" [t]. Dans ce cas, le mécanisme de flux d'air est omis.

Certaines paires de consonnes comme p :: b, t :: d sont parfois appelées fortis et lenis , mais il s'agit d'une distinction phonologique plutôt que phonétique.

Quelle utilité de connaître cette différence ? Deux exemples : Une consonne sonore rallonge la voyelle qui précède. Le /i/ de lid est plus long que le /i/ de lip. Le "ed" du prétérit se prononce sourd /t/ après une sourde et sonore /d/ après une sonore.

2.2.2 Les Voyelles

En phonétique, on appelle voyelle un son du langage humain dont le mode de production est caractérisé par le libre passage de l'air dans les cavités situées au-dessus de la glotte, à savoir la cavité buccale et/ou les fosses nasales. Ces cavités servent de filtres dont la forme et la contribution relative à l'écoulement de l'air influent sur la qualité du son obtenu. La plupart des voyelles utilisées dans les langues sont sonores, c'est-à-dire qu'elles sont prononcées avec une vibration des cordes vocales, le chuchotement utilise – par définition des voyelles sourdes.

Son dans lequel le flux d'air des poumons passe par la bouche, qui fonctionne comme une chambre de résonance , avec une obstruction minimale et sans frottement audible; par exemple, le i en «forme» et le a dans le «paquet». Bien qu'ils soient généralement produits avec des cordes vocales vibrantes, les voyelles peuvent être prononcées sans cette vibration, ce qui produit un son sans voix ou chuchoté. Du point de vue de la phonétique articulatoire .Une voyelle est produite par l'action coordonnée des cordes vocales et de différents articulateurs qui se situent entre le larynx et les lèvres : la langue, le voile du palais, les dents, les lèvres. Elles sont classées en fonction de la position de la langue et les lèvres et, parfois, selon que l'air est ou non libéré par le nez.

Une voyelle haute (telle que i dans «machine» et u dans «rule») est prononcée avec la langue cambrée vers le toit de la bouche. Une voyelle basse (comme un en « father» ou « had ») est produit avec la langue relativement plat et bas dans la bouche et la bouche ouverte un peu plus large que pour les voyelles élevées. Les voyelles moyennes (telles que e dans «bed» et o dans «pole») ont une position de la langue entre les extrêmes haute et basse.

Les voyelles hautes, moyennes et basses sont également classées selon une dimension d'avant en arrière. Une voyelle antérieure est prononcée avec la partie la plus haute de la langue poussée en avant dans la bouche et légèrement cambrée. Le a dans «had», le e dans «bed» et le i dans «fit» sont des voyelles antérieures. Une voyelle arrière - par exemple, le u dans la «rule» et le o dans le

«pole» sont produits avec la partie arrière de la langue relevée vers le palais mou (velum).

La forme et la position des lèvres donnent une troisième dimension articulatoire selon laquelle les voyelles sont classées. Les lèvres peuvent être arrondies ou étendues, dans ce qu'on appelle la labialisation.

Les caractéristiques articulatoires supplémentaires décrivant l'articulation des voyelles sont «large» et «étroit», «tendu» (fortis) et «laxiste» (lenis).

Large et étroit se réfèrent à la position de la langue. Pour former une voyelle étroite, la racine de la langue est rétractée vers la paroi du pharynx et le pharynx est rétréci. Pour former une voyelle large, la racine de la langue est avancée de sorte que le pharynx est élargi.

Tension et relâchement sont des termes moins clairement définis. Les voyelles tendues sont articulées avec un effort musculaire plus important, des positions de la langue légèrement plus élevées et des durées plus longues voyelles laxistes.

Toutes les voyelles peuvent être divisées en deux catégories principales: diphtongues et monophthongs. Les diphtongues sont des voyelles glissantes dans l'articulation desquelles se produit une transition continue d'une position à une autre. À cet égard, les diphtongues doivent être comparées aux voyelles dites pures, oules monophthongs, c'est -à-dire les voyelles immuables ou stables. Bien qu'il s'agisse de sons simples, les diphtongues sont généralement représentées, dans une transcription phonétique de la parole, au moyen d'une paire de caractères indiquant les configurations initiale et finale de l'appareil vocal. Un grand nombre des voyelles dans la plupart des dialectes de l'anglais sont des diphtongues - par exemple, les voyelles de «out» et de «ice».

Les semi - voyelles sont des sons produits de la même manière que les voyelles mais sont utilisés et perçus comme des consonnes. Les exemples incluent le y dans «yawn» et le w dans «walk». (20)

2.2.3 Transcription Orthographique Phonétique (21)

La transcription, au sens linguistique, est la représentation systématique du langage sous forme écrite. Certains linguistes considèrent que la seule base de la transcription doit être l'énoncé, même si des textes déjà existants dans un autre système d'écriture peuvent également servir de support. Elle joue également un rôle important pour plusieurs sous-domaines de la technologie vocale. L'exemple typique de transcription est l'enregistrement écrit par un greffier d'un compte-rendu d'une audience devant un tribunal, par exemple en cas de procédure pénale. Cet article se concentre sur la transcription dans le domaine de la linguistique.

De manière générale, il existe deux approches possibles à la transcription linguistique : **La transcription phonétique** se concentre sur les propriétés phonétiques et phonologiques du langage parlé. Les systèmes de transcription phonétique fournissent ainsi les règles de représentation des sons individuels ou des phonèmes en symboles écrits.

Les systèmes de **transcription orthographique**, au contraire, définissent les règles de représentation des mots exprimés sous leur forme orale sous la forme écrite d'un langage donné, tout en respectant ses règles d'orthographe. La transcription phonétique utilise des ensembles de caractères expressément définis, dans la plupart des cas l'Alphabet Phonétique International (API).

Le type de transcription choisi dépend majoritairement des intérêts de recherches poursuivis.

Étant donné que la transcription phonétique met en avant la nature phonétique du langage, elle est plus utile dans le cadre d'analyses phonologiques ou phonétiques. En revanche, la transcription orthographique dispose d'un élément lexical et morphologique aux côtés de l'élément phonétique (l'aspect représenté, et à quel point celui-ci est représenté, dépend du langage en question et de son orthographe). Elle est ainsi plus commode à utiliser en cas d'analyses qui étudient les aspects liés au sens d'un langage parlé. La transcription phonétique est sans aucun doute plus systématique au sens scientifique, mais est également plus difficile à assimiler,

demande plus de temps à réaliser, et est moins aisée à mettre en pratique que la transcription orthographique [7].

2.3 Système de segmentation de la parole

Le signal de la parole peut être représenté par deux ensembles, les composantes analogiques qui nous intéressent et le reste est considéré comme du bruit.

Le schéma proposé est testé sur un certain nombre de données de parole enregistrées. On constate que la méthode est capable d'extraire les composantes de la parole (c'est-à-dire les intervalles actifs) de composants non liés à la parole (c'est-à-dire intervalles inactifs).

La segmentation des composants du signal implique l'exploitation des paramètres temporels, spectraux et spatiaux. Dans la littérature, deux techniques principales ont été employées : la segmentation implicite et explicite. La première méthode sépare l'énoncé de parole en segments sans information explicite comme la transcription phonétique ; elle définit le segment implicitement en se basant sur la stabilité des parties spectrales du signal. L'autre méthode sépare les énoncés entrants en segments qui sont explicitement définies par la transcription phonétique. (22)

Plusieurs méthodes ont été proposées pour la segmentation automatique de parole en phonèmes.

Dans cette méthode, la segmentation est obtenue en estimant le meilleur chemin dans une séquence de parole à l'aide de la programmation dynamique HMM (Hidden Markov Model). En outre, un algorithme basé sur la détection du pic des MFCC a été proposé.

Ces approches sont basées sur l'information textuelle incluse dans le signal de parole. Elle est utile pour certaines applications telles que l'étiquetage automatique des corpus de parole. Cependant, pour des applications tel que la reconnaissance de parole en temps réel, la traduction automatique et l'apprentissage assisté par ordinateur, l'information textuelle est indisponible.

2.3.1 Modèle générale

Le modèle général du système de segmentation automatique de la parole comprend la préparation des données, la modélisation du langage, la construction du modèle HMM, la segmentation HMM et les sous-systèmes de vérification des données.

Pour construire notre système de segmentation automatique du corpus Anglais en phonème, nous avons implémenté comme technique: les Modèles de Markov Cachés ou HMM. Leurs capacités de segmenter entre les différentes unités acoustiques montrent l'efficacité et la nécessité de les utiliser pour résoudre le problème de la segmentation en phonème.

En outre, un algorithme basé sur la détection du pic des MFCC a été proposé. On utilise aussi à l'entrée un fichier wav et un fichier de transcription puisque à l'entrée HMM pour voir la composition des phrases. Ces approches sont basées sur l'information textuelle incluse dans le signal de parole.

La sortie du système représente un fichier test contenant les banderais déjà connu.

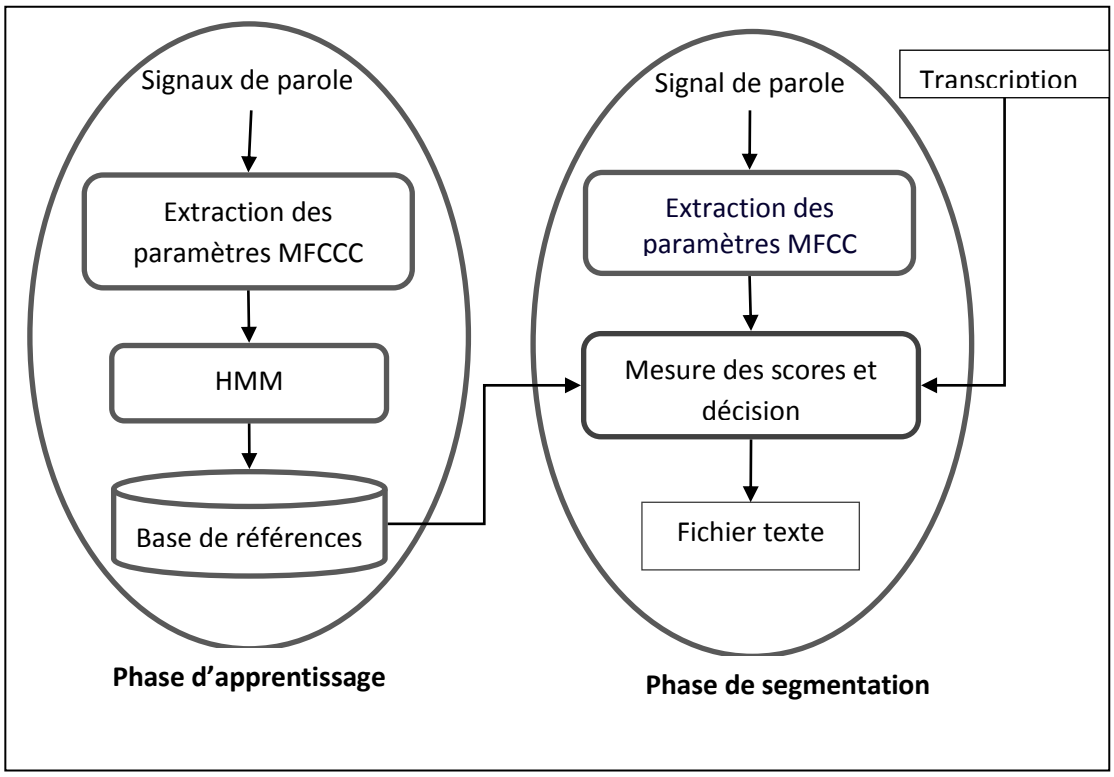


Figure 2.1 : Segmentation automatique des signaux de parole

2.3.2 Extraction des paramètres

Bien souvent, le message utile est noyé parmi d'autres informations : en compréhension de la parole par exemple, l'identité du locuteur est souvent non pertinente. Une question fondamentale se pose quant aux critères précisant la pertinence des paramètres à extraire dans le signal.

L'étude des coefficients MFCC du signal permet d'extraire des caractéristiques de celui-ci autour de la FFT et de la DCT, convertis sur une échelle de Mel. L'échelle de Mel est une échelle psycho-acoustique de hauteurs des sons, au sens de leur repérage entre grave et aigu, dont l'unité est le Mel.

Le Mel est relié au hertz (Hz), l'unité de mesure du Système international pour les fréquences, par une relation établie par des expériences basées sur l'audition humaine.

Il s'agit de la Méthode la plus utilisée pour représenter un signal en reconnaissance de la parole, car très robuste. Son principal avantage est que les coefficients obtenus sont décarrelles.

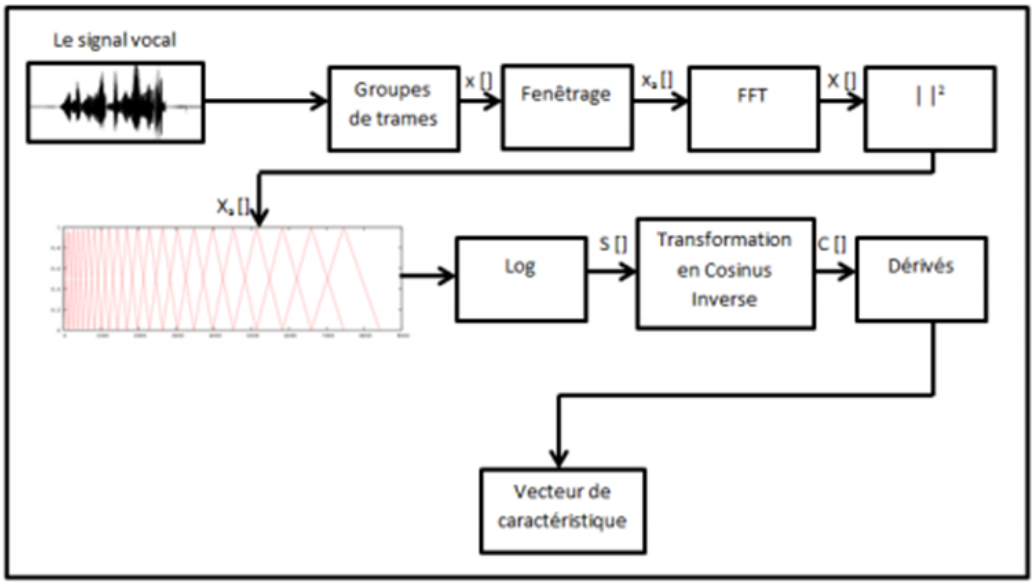


Figure 2.2 : Etapes de calcul d'un vecteur caractéristique de type MFCC

Le calcul des coefficients MFCC est réalisé de la manière suivante (5) :

- groupement de trames : Le signal acoustique continu est segmenté en trames de N échantillons, avec un pas d'avancement de M trames (M < N), c'est-à-dire que deux trames consécutives se chevauchent sur (N – M) échantillons. Les valeurs couramment utilisées pour M et N sont respectivement 10 et 20. Comme prétraitement, il est d'usage de procéder à la préaccentuation du signal Préaccentuation du signal, il s'agit de faire ressortir les hautes fréquences avec un filtre passe-haut de la forme :

$$H(z) = 1 - 0.9z^{-1} \quad (4)$$

- Découpage du signal en fenêtre : Si nous définissons $w(n)$ comme fenêtre où $0 < n < N - 1$ et N représente le nombre d'échantillons dans chacune des trames, alors le résultat du fenêtrage est le signal x_a, donné par la formule.

$$x_a = x(n)w(n), \qquad 0 < n < N-1 \quad (5)$$

Les fenêtres les plus utilisées sont : Fenêtre de Hamming, Fenêtre rectangulaire, Fenêtre triangulaire, Fenêtre de Hann, Fenêtre de Blackman.

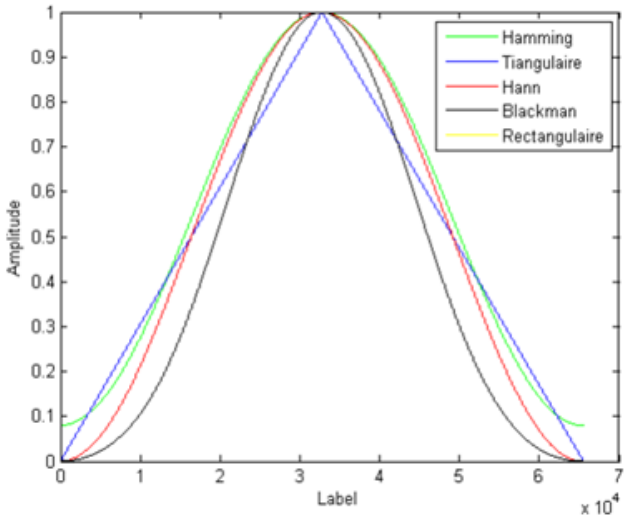


Figure 2.3 : Les fenêtres les plus utilisées

- Application d'une transformée de Fourrier sur chacune des portions: Au cours de cette étape chacune des trames, de N valeurs, est convertie du domaine temporel au domaine fréquentiel. La FFT est un algorithme rapide pour le

calcul de la transformée de Fourier discret (DFT) et est définie par la formule en dessous. Les valeurs obtenues sont appelées le spectre.

$$\mathrm{x[k]} = \sum_{n=0}^{N-1} x_a[\mathrm{n}] e^{\frac{-2\mathrm{j}\pi\,\mathrm{kn}}{\mathrm{N}}}, \quad 0 \leq \mathrm{k} \leq \mathrm{N} - 1 \tag{6}$$

En général, les valeurs $X[k]$ sont des nombres complexes et nous nous utilisons que leurs valeurs absolues (énergie de la fréquence).

- Filtrage sur l'échelle Mel, Le spectre d'amplitude est pondéré par un banc de M filtres triangulaires espacés selon l'échelle Mel. Dans l'échelle de mesure Mel, la correspondance est approximativement linéaire sur les fréquences au-dessous de 1kHz et logarithmique sur les fréquences supérieures à celle-ci. Cette relation est donnée par la formule:

$$m = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

- Calcul du cepstre sur l'échelle Mel : Le cepstre sur l'échelle de fréquence Mel est obtenu par le calcul de la transformée en cosinus discrète DCT (Discrète Cosinus Transform) du logarithme de la sortie des M filtres (reconversion du log-Mel-spectre vers le domaine temporel).

- Calcul des caractéristiques dynamiques des MFCC : Les changements temporels dans le cepstre (C) jouent un rôle important dans la perception humaine et c'est à travers les dérivées des coefficients (Δ, coefficients delta ou vélocité) et les dérivées secondes (ΔΔ, coefficients delta du second ordre ou accélération) des MFFC statiques que nous pouvons mesurer ces changements.En résumé, un système de parole typique de l'état de l'art effectue premièrement un échantillonnage à une fréquence de 16 kHz et extrait les traits suivants :

$$\begin{pmatrix} C_k \\ \Delta C_k \\ \Delta\Delta C_k \end{pmatrix} \tag{8}$$

Ou` :

- C_k : vecteur MFCC de la k^{ime} trame

- ΔC_k : dérivée première des MFCCs
- $\Delta\Delta C_k$: seconde dérivée des MFCCs [8].

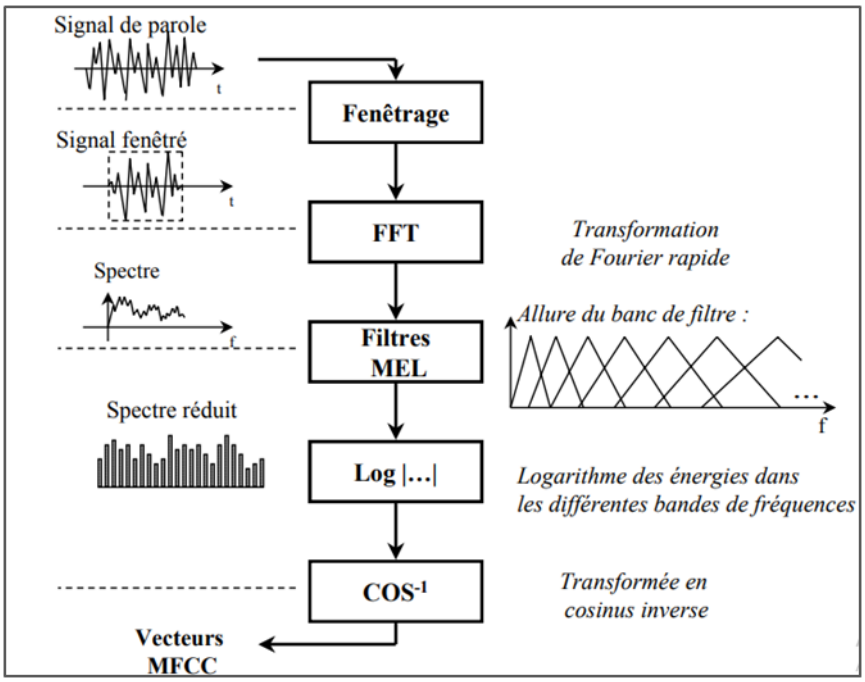


Figure 2.4 : Calcul des MFCC avec les signaux

2.3.3 Modèles de Markov Cachés (HMM)

Le Modèle de Markov Caché (Hidden Markov Model) est une méthode statistique puissante pour caractériser les échantillons de données observés d'un processus à temps discret. Elle apporte non seulement un moyen efficace de construction de modèles paramétriques, mais elle incorpore aussi le principe de programmation dynamique pour unifier la segmentation et la classification de séquence de données variant dans le temps.

Les modèles de Markov cachés peuvent être utilisés pour représenter la séquence de sons dans une section d'unités de parole telles que les phonèmes. Phonème, un son élémentaire de la parole, peut être modélisé par un individu HMM de gauche à droite.

Dans la modélisation d'un processus par un HMM, les échantillons peuvent être caractérisés par un processus paramétrique aléatoire dont les paramètres peuvent être estimés dans un cadre de travail bien défini.

Les HMM sont devenus la méthode la plus couramment utilisée pour la modélisation des signaux de parole dans les applications suivantes : reconnaissance automatique de la parole, suivi de la fréquence fondamentale et des formants, synthèse vocale, traduction automatique, étiquetage syntaxique, compréhension du langage oral, traduction automatique...

Un processus stochastique est markovien si son évolution est entièrement déterminée par une probabilité initiale et des probabilités de transitions entre états (son évolution ne dépend pas de son passé mais uniquement de son états présent, l'état courant du système contient toute l'information pour prédire son état futur. (5)

Ci-dessous, nous présentons les trois problèmes de base à résoudre pour l'application de cette méthode (23):

- Le problème d'évaluation : Quelle est la probabilité d'un modèle générant une séquence d'observation ? Ce problème est résolu par l'application de l'algorithme FORWARD.
- Le problème de décodage : Quelle est la séquence d'états la plus probable pour un modèle et une séquence d'observation donnés ? On utilise l'algorithme VITERBI pour effectuer cette tâche.
- Le problème d'apprentissage : Comment peut-on ajuster les paramètres du modèle pour maximiser la vraisemblance (probabilité jointe) de génération d'une séquence d'observation ? Les algorithmes de BAUM-WELCH et de VITERBI permettent d'effectuer l'apprentissage.

Dans les applications de la parole, on utilise fréquemment les HMM continus, où l'observation n'appartient pas à un ensemble discret mais à une distribution (le plus souvent normale). Ainsi, une topologie gauche-droite pour un HMM continu permet de modéliser les états successifs d'un phonème pour un signal de parole. Plus généralement, l'objectif à atteindre est la détermination à partir de vecteurs acoustiques de la séquence phonétique prononcée.

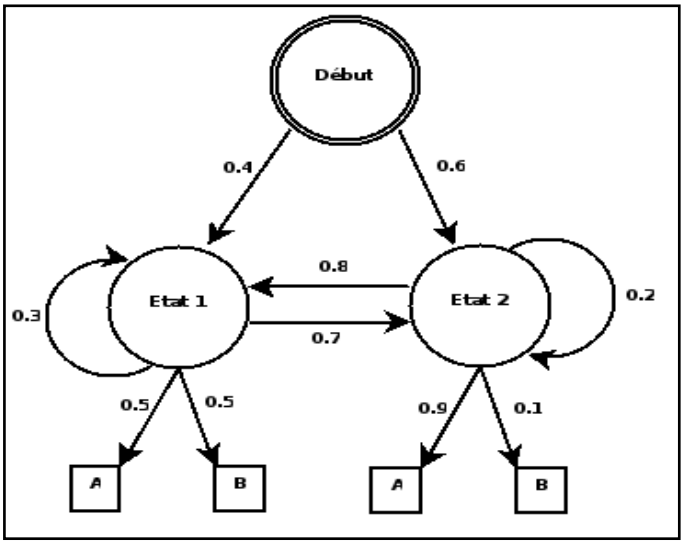


Figure 2.5 : Exemple d'un HMM à 2 états

- **Les états** : un instant donnée, la description du système est donné par un état donné
- **Les transitions** : ce sont le changement d'états

Un modèle markovien d'ordre k considère que l'état du système à un instant t ne dépend que de l'état aux k instants précédents. Un HMM est noté $\Lambda = (A, B, \pi)$ et se définit par :

- Ses états, en nombre n, qui composent l'ensemble$S = \{s_1, s_2, \ldots, s_n\}$. L'état où se trouve le HMM à l'instant t est noté $\mathrm{q_t}$ $(\mathrm{q_t} \in S)$;
- M symboles observables dans chaque état. L'ensemble des observations possibles (l'alphabet) est noté $V = \{v_1, v_2, \ldots, v_M\}$.

  $O_t \in V$ est le symbole observé à l'instant t .
- Une matrice A de probabilités de transition entre les états : a_{ij} représente la probabilité que le modèle évolue de l'état i vers l'état j :

$$a_{ij} = A(i,j) = P\left(q_{t+1} = s_j \mid q_t = s_i\right) \tag{9}$$

Avec :

$$a_{i,j} \geq 0 \quad \forall_{i,j} \text{ et } : \sum_{j=1}^{n} a_{i,j} = 1 \tag{10}$$

- Une matrice B de probabilités d'observation des symboles dans chacun des états du modèle : $b_j(k)$ représente la probabilité que l'on observe le symbole v_k alors que le modèle se trouve dans l'état j, soit :

$$b_j\,(\mathrm{k}) \;=\; \mathrm{P}(O^t \;=\; v_k \mid q_t \;=\; s_j\,) \;\; 1 \;\leq\; \mathrm{j} \;\leq\; \mathrm{n}, 1 \;\leq\; \mathrm{k} \;\leq\; \mathrm{M} \qquad (11)$$

Avec

$$\mathrm{bj}\,(\mathrm{k}) \geq\; 0\;\forall \mathrm{j}, \mathrm{k} \;\; \text{et} \;\; : \sum_{k=1}^{M} b_j(k) = 1 \qquad (12)$$

Un vecteur π de probabilités initiales : $\pi \;=\; \{\pi_i\},\; i = 1,2,\ldots,n$. Pour tout état i, π_i est la probabilité que l'état de départ du HMM soit l'état i :

$$\pi_i \;=\; \mathrm{P}(q_1 \;=\; s_i)\; 1 \;\leq\; \mathrm{i} \;\leq\; \mathrm{n} \qquad (13)$$

Avec

$$\pi_i \;\geq\; 0 \;\; \forall \mathrm{i} \; \text{et} : \sum_{i=1}^{n} \pi_i = 1 \qquad (14)$$

- Un ou plusieurs états finals. Ici, nous supposons pour simplifier que le processus peut s'arrêter dans n'importe quel état, autrement dit que tout état est final.

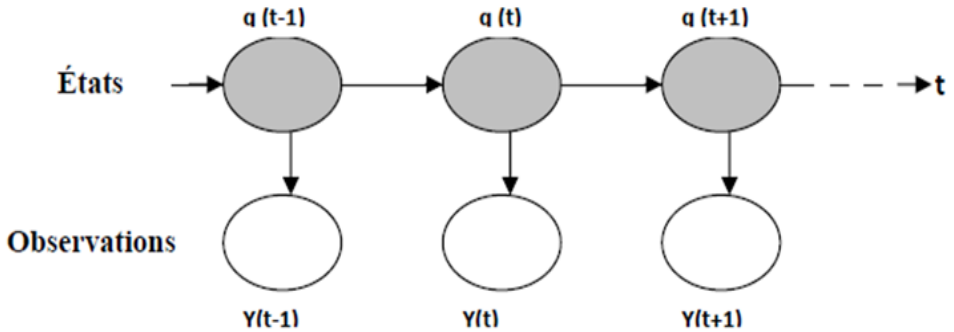


Figure 2.6 : Représentation d'un HMM

Pour résumer, un HMM comprend deux paramètres constants, N et M, représentant le nombre total d'états et de la taille des alphabets d'observation. Il est aussi caractérisé par les observations alphabets O et les trois matrices de probabilité

A, B et π. Généralement, on peut représenter un HMM par la notation suivante : $\Lambda = (A, B, \pi)$. (22)

a Phase d'apprentissage

Afin d'effectuer la reconnaissance, il faut avoir un modèle de la séquence à reconnaitre que l'on pourra ensuite comparer aux séquences inconnues. Pour construire ce modèle, on pourrait utiliser les connaissances a priori dont on dispose sur le système.

Celles-ci sont généralement insuffisantes pour donner des résultats convaincants. Nous allons donc plutôt faire appel à une méthode d'apprentissage statistique. Celle-ci va permettre de modifier les probabilités des différentes transitions du HMM afin de le rapprocher du modèle recherché.

Etant donné de Markov caché d'architecture fixée, l'apprentissage vise à déterminer ses paramètres (matrice de probabilités de transitions, matrice de probabilité d'émission et matrice de probabilités initiales) (24). Cet apprentissage se fait par une approche rigoureuse qui consiste à chercher les paramètres λ Δ qui maximisent :

$$P(O|\Lambda) = \prod_{k=1}^{k} P(Y^k \mid \Lambda) \tag{15}$$

Il faut en effet que Λ ait une probabilité maximale d'émettre les séquences d'apprentissage. Cet entrainement, qui suit le principe du maximum de vraisemblance, s'effectue suivant l'algorithme d'entrainement de **Baum-Welch** (L'algorithme de Baum-Welch est un cas particulier de l'algorithme espérance-maximisation et utilise l'algorithme Forward-Backward. Il permet de ré-estimer les paramètres de manière itérative.) (24)

Principe (24) **:** Supposons disposer d'un ensemble de séquences $O = \{O^1, \cdots, O^m\}$, dont l'élément courant est noté O_k. Le but de l'apprentissage est de déterminer les paramètres d'un HMM d'architecture fixée $\Lambda = (A, B, \pi)$, qui maximisent la probabilité $P(O|\Lambda)$**.** Comme on suppose les séquences d'apprentissages tirées indépendamment, on cherche donc à maximiser :

$$P(\mathrm{O}|\Lambda) = \prod_{k=1}^{m} P\left(O^k \ |\Lambda\right) \tag{16}$$

L'idée est d'utiliser une procédure de ré-estimation qui affine le modèle petit à petit selon les étapes suivantes :

- choisir un ensemble initial Λ_0 de paramètres ;
- calculer Λ_1 à partir de Λ_0, puis Λ_2 à partir de Λ_1, etc.
- répéter ce processus jusqu'à un critère de fin.

  Pour chaque étape p d'apprentissage, on dispose de Λ_p et on cherche un Λ_{p+1} qui doit vérifier

$$P\left(O^k \mid \Lambda_{p+1}\right) \geq P\left(O^k \mid \Lambda_p\right) \tag{17}$$

  Soit :

$$\prod_{k=1}^{m} P\left(O^k \mid \Lambda_{p+1}\right) \geq \prod_{k=1}^{m} P\left(O^k \mid \Lambda_p\right) \tag{18}$$

Λ_{p+1} doit donc améliorer la probabilité de l'émission des observations de l'ensemble d'apprentissage. La technique pour calculer Λ_{p+1} à partir de Λ_p consiste à utiliser l'algorithme EM. Pour cela, on effectue un comptage de l'utilisation des transitions A et des distributions B et π du modèle Λ_p quand il produit l'ensemble O. Si cet ensemble est assez important, ces fréquences fournissent de bonnes approximations a posteriori des distributions de probabilités A, B et π et sont utilisables alors comme paramètres du modèle Λ_{p+1} pour l'itération suivante.

La méthode d'apprentissage EM consiste donc dans ce cas à regarder comment se comporte le modèle défini par Λ_p sur O, à ré-estimer ses paramètres à partir des mesures prises sur O, puis à recommencer cette ré estimation jusqu'à obtenir une convergence.

Dans les calculs qui suivent, on verra apparaître en indice supérieur la lettre k quand il faudra faire référence à la séquence d'apprentissage concernée. L'indice p, qui compte les passes d'apprentissage, sera omis : on partira d'un modèle noté simplement Λ et on calculera celui qui s'en déduit.

Les formules de ré estimation : On définit $\xi_t^k(\mathrm{i},\mathrm{j})$ comme la probabilité, étant donnés une phrase O^k et un HMM Λ, que ce soit l'état s_i qui ait émis la lettre de rang t de O^k et l'état **s_j** qui ait émis celle de rang $t+1$. Donc :

$$\xi_t^k(\mathrm{i},\mathrm{j}) = \mathrm{P}(q_t = \mathrm{si}, q_{t+1} = \mathrm{sj} \mid O^k, \Lambda) \tag{19}$$

Ce qui se récrit :

$$\xi_t^k(\mathrm{i},\mathrm{j}) = \frac{\mathrm{P}(q_t = \mathrm{s_i}, q_{t+1} = \mathrm{s_j}, O^k \mid \Lambda)}{\mathrm{P}(O^k \mid \Lambda)} \tag{20}$$

Par définition des fonctions **forward-backward,** on en déduit :

$$\xi_t^k(\mathrm{i},\mathrm{j}) = \frac{\alpha_t^k(\mathrm{i}) a_{i,j}\, b_j\left(O_{t+1}^k\right) \beta_{t+1}^k(\mathrm{j})}{\mathrm{P}(O^k \mid \Lambda)} \tag{21}$$

On définit aussi la quantité $\gamma_t^k(\boldsymbol{i})$ comme la probabilité que la lettre de rang t de la phrase O_k soit émise par l'état s_j**.**

$$\gamma_t^k(\mathrm{i}) = \mathrm{P}(\mathrm{qt} = \mathrm{s_i} \mid O^k, \Lambda) \tag{22}$$

Soit :

$$\xi_t^k(\mathrm{i},\mathrm{j}) = \frac{\alpha_t^k(\mathrm{i}) a_{i,j}\, b_j\left(O_{t+1}^k\right) \beta_{t+1}^k(\mathrm{j})}{\mathrm{P}(O^k \mid \Lambda)} \tag{23}$$

$$\gamma_t^k(\mathrm{i}) = \sum_{j=1}^{n} P\left(\mathrm{q_t} = \mathrm{s_i}, \mathrm{q_{t+1}} = \mathrm{s_j} \mid O^k, \Lambda\right) = \frac{\sum_{j=1}^{n} \mathrm{P}(\mathrm{q_t} = \mathrm{s_i}, \mathrm{q_{t+1}} = \mathrm{s_j}, O^k \mid \Lambda)}{\mathrm{P}(O^k \mid \Lambda)} \tag{24}$$

On a la relation :

$$\gamma_t^k(\mathrm{i}) = \sum_{j=1}^{n} \xi_t^k(\mathrm{i},\mathrm{j}) \frac{\alpha_t^k(\mathrm{i}) \beta_t^k(\mathrm{i})}{\mathrm{P}(O^k \mid \Lambda)} \tag{25}$$

Le nouveau modèle HMM se calcule à partir de l'ancien en ré estimant π, A et B par comptage sur la base d'apprentissage. On mesure les fréquences :

$$\pi_\mathrm{i} = \frac{1}{m} \sum_{k=1}^{m} \gamma_1^k(\mathrm{i}) \tag{26}$$

$$a_{ij} = \frac{\sum_{k=1}^{m} 1 \sum_{t=1}^{|O_k|-1} \xi_t^k (i,j)}{\sum_{k=1}^{N} 1 \sum_{t=1}^{|O_k|-1} \gamma_t^k (i)} \quad (27)$$

$$b_j(l) = \frac{\sum_{k=1}^{m} 1 \sum_{\substack{t=1 \\ avec\ Ok\ t = vl}}^{|O_k|-1} \gamma_t^k (j)}{\sum_{k=1}^{m} 1 \sum_{t=1}^{|O_k|-1} \gamma_t^k (j)} \quad (28)$$

Ces formules ont été établies par Baum, comme une application de la procédure EM (*Expectation-Maximisation)* à l'apprentissage des paramètres HMM. La suite des modèles construits par l'algorithme de Baum-Welch vérifie la relation cherchée :

$$P(O|\Lambda_{p+1}) \geq P(O|\Lambda_p) \quad (29)$$

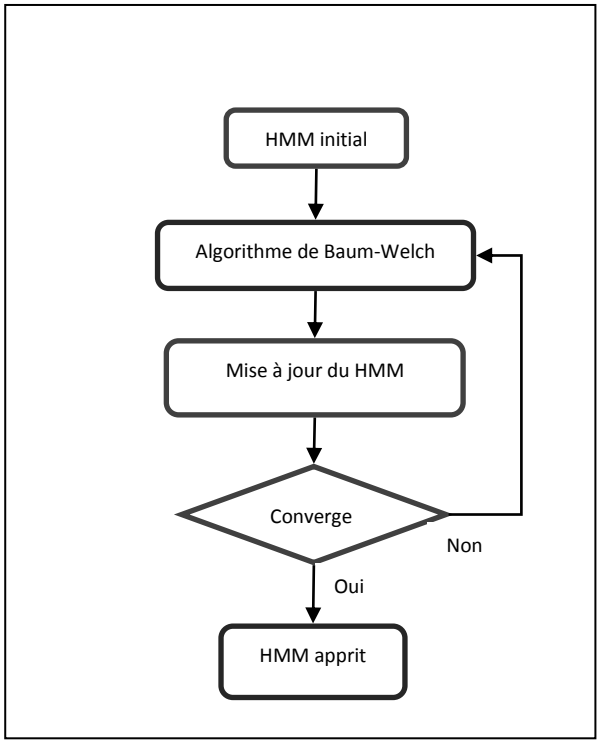


Figure 2.7 : Processus d'apprentissage

Remarque :

- Le choix du modèle initial influe sur les résultats ; par exemple, si certaines valeurs de A et B sont égales à 0 au départ, elles le resteront jusqu'à la fin de l'apprentissage. Ceci permet en particulier de garder la structure dans les modèles gauches-droits.

- L'algorithme converge vers des valeurs de paramètres qui assurent un maximum local de $P(O|\Lambda)$. Il est donc important, si l'on veut être aussi près que possible du minimum global, de bien choisir la structure et l'initialisation.
- Le nombre d'itérations est fixé empiriquement. L'expérience prouve que, si le point précédent a été correctement traité, la stabilisation des paramètres ne correspond pas à un sur-apprentissage (24) : il n'y a donc en général pas besoin de contrôler la convergence par un ensemble de validation. Mais cette possibilité est évidemment toujours à disposition.

b Phase de segmentation

Pour la phase de segmentation, on utilise l'algorithme de Viterbi.

Plusieurs méthodes existent pour trouver le chemin le plus probable des états cachés, mais la plus commune est l'algorithme de Viterbi.

L'algorithme Viterbi est un algorithme issu de la programmation dynamique permettant de déterminer un chemin le plus probable associé à une séquence d'émissions $e_1 \dots e_k$ pour un HMM. Cet algorithme se base sur un calcul itératif pour chaque état caché S_1 du chemin le plus probable atteignant cet état à l'étape i de l'exécution d'un processus Markovien. On note par la suite cette probabilité $p(e_1 \dots e_i)_{S_1}$. Cette itération est effectuée chaque à étape de l'exécution du processus Markovien. L'algorithme est donc de complexité linéaire par rapport à la taille de la séquence d'émissions.

La probabilité du chemin le plus probable atteignant l'état s_1 à l'étape i peut être calculé récursivement selon la formule suivante :

$$p(e_1 \dots . e_\iota) = \max_{s_k \epsilon S}(p(e_1 \dots . . e_{i-1})_{sk} \ . p(s_k; s_\iota) \, . p(\, s_\iota; e_i)) \tag{30}$$

Si l'état caché S_m permet de maximiser $p(e_1 \dots . . e_{i-1})_{sk} \ . p(s_k; s_\iota) \, . p(\, s_\iota; e_i)$, le chemin le plus probable atteignant s_l à l'étape $(i-1)$ est composé du chemin le plus probable atteignant S_m à l'étape $(i-1)$ et de la transition entre S_m et S_l.

L'algorithme Viterbi itère de e_i à $e_1 \ldots e_n$. Cet algorithme calcule pour chaque état caché du modèle $s_l \in S$ le chemin le plus probable atteignant cet état à l'étape i et la probabilité de ce chemin partiel. Pendant l'itération, les différents chemins ainsi que leurs probabilités respectives sont stockés respectivement dans la structure de données Paths ou Probas. Cette structure de données associe à chaque état caché le chemin le plus probable atteignant cet état ou la probabilité de celui-ci. L'´équation précédente permet de déduire la valeur de ce chemin ainsi que sa probabilité au regard des valeurs calculées à l'étape précédente de l'itération (25).

Afin de résoudre le problème de décodage, l'algorithme de Viterbi est employé.

Le critère d'optimalité ici est de rechercher un meilleur ordre simple d'état par la technique modifiée de la programmation dynamique. L'algorithme de Viterbi est un algorithme de recherche parallèle, à savoir il recherche le meilleur ordre d'état en traitant tous les états en parallèle. Nous devons maximiser $P\ (Q|O,\wedge)$ pour détecter le meilleur ordre d'état. Soie la probabilité $\delta_t(i)$ qui représente la probabilité maximale le long du meilleur chemin probable d'ordre d'état d'une séquence d'observation donné après *t* instants et en étant à l'état *i* ;

$$\delta_t(i) = \max_{sk\ S} P\ [q_1, q_2 \ldots\ldots q_{t-1}\ , q_t\ =\ S_i, o_1 \ldots o_t |\ \wedge] \tag{31}$$

La meilleure séquence d'états et retournée par une autre fonction $\psi_t(j)$. Cette fonction tient l'index de l'instant $t-1$, à partir duquel la meilleure transition est faite à l'état actuel. L'algorithme complet est comme suit (5) :

Initialisation : $\psi_1(i)$=0 ; $\delta_1(i) = \pi(i)P(o_1|i\ )$;

Induction :

$$\delta_t(i) = \max_{i' \in S}\big(\delta_{t-1}(\ i')\ P(i' \longrightarrow i)\big)\ P(o_t|i\ )$$

$$\psi_t(i) = \arg\max_{i' \in S}(\delta_{t-1}(i')P(i' \longrightarrow i))$$

2.4 Conclusion

Dans ce travail, nous avons présenté les différentes étapes pour mettre en œuvre le système de segmentation automatique de la parole en phonèmes.

L'extraction des meilleurs paramètres aide, sans aucun doute, dans notre traitement. L'intelligence artificielle peut intervenir pour trouver les paramètres pertinents ou utiliser n'importe quels représentants de la parole pour faire la segmentation ou la classification.

Les Modèles de Markov cachés *HMM*, précisément les Chaînes de Markov Cachées, sont donc des outils de modélisation très utiles. Leur choix en vue de résoudre un problème bien défini nécessite une spécification complète de leurs éléments ; à savoir: la matrice de transition, les distributions initiales et les densités d'émission des observations.

Dans le chapitre suivant, nous allons voir que deux phases principales sont réalisées: la phase d'apprentissage et la phase de reconnaissance. La phase d'apprentissage est mise en œuvre à l'aide de différents modules de l'outil HTK. Les résultats obtenus démontrent la puissance de l'utilisation de HTK pour résoudre le problème de la segmentation de la parole. Cependant, même avec ce système, la segmentation finale doit être vérifiée par l'utilisateur.

Chapitre 3 Implémentation et Evaluation Expérimentale

3.1 Introduction

Ce chapitre illustre les faits des chapitres précèdent par la mise en test des approches considérées. Nous présentons différents résultats expérimentaux obtenu sur la base TIMIT et nous concluons sur la performance de notre système.

Nous allons implémenter le système de la segmentation automatique de parole en phonème, notre système est mis en œuvre à l'aide de HMM sous environnement Matlab. Dans cette méthode, la segmentation est obtenue en estimant le meilleur chemin dans une séquence de parole à l'aide de la programmation dynamique ou les HMM. En outre, un algorithme basé sur la détection du pic des MFCC a été proposé pour la segmentation automatique de la parole. Ces approches sont basées sur l'information textuelle incluse dans le signal de parole. Elle est utile pour certaine application telle que l'étiquetage automatique des corpus de parole.

Nous avons implémenté plusieurs configurations suivant le nombre d'états du HMM, et le nombre des coefficients MFCC, finalement, nous avons estimé les performances en calculant le Taux de Segmentation Correct en fonction du nombre de Composantes Gaussiennes CG, (pour modéliser la sortie du HMM) (22).

3.2 Matériels et méthode utilisée

Nos recherches se sont orientées vers l'implémentation SAPh dans l'environnement HTK. C'est un outil pour construire et manipuler des modèles de Markov cachés. HTK est principalement utilisé pour la reconnaissance automatique de

la parole, bien qu'il ait été utilisé pour de nombreuses autres applications, notamment la synthèse de la parole, reconnaissance des caractères et plusieurs chercheurs du monde entier.

Pour effectuer la tâche de segmentation automatique de la parole. La préparation des données, initialisation reconnaissance et la segmentation en sont les principales étapes.

Dans la segmentation de la parole basée sur un corpus, il est nécessaire de préparer et de tester l'ensemble de données. À cette fin, une technique d'échantillonnage aléatoire systématique consiste à scinder le corpus de texte et de parole en ensembles d'apprentissage (Data Training) et ensembles de test (Data Test). Training Data est utilisé à des fins de modélisation linguistique et acoustique, alors que Data Test est utilisé pour l'évaluation de la segmentation automatique HMM. Étant donné que HTK n'utilise pas les données de parole, les transcriptions sont effectuées, et leur paramétrage est également requis dans le cadre de la préparation des données. Le paramétrage des données de parole s'effectue via un processus d'extraction de caractéristiques(MFCC).

Après avoir obtenu les résultats de la modélisation linguistique et de la segmentation HMM affecte la lettre ou le phonème correspondant au signal acoustique correspondant à la formation et à la prononciation optimale du mot sélectionné. À la fin, les phonèmes avec leurs limites de temps sont obtenus en tant que sortie de la segmentation de parole automatique basé sur HMM.

3.2.1 Corpus

Pour le développement de notre système de reconnaissance, nous avons utilisé la base de données acoustique américaine TIMIT pour plusieurs raisons. Tout d'abord, cette base a été constituée pour illustrer au mieux la variabilité acoustique de l'anglais américain, et elle est fournie avec une segmentation phonétique de référence qui simplifie l'apprentissage initial des modèles phonétiques. De plus, TIMIT peut être considérée comme une base de données de référence. Sa large diffusion dans la

communauté internationale permet une évaluation objective des performances des systèmes développés.

Le corpus TIMIT de lecture de la parole est conçu pour fournir des données de parole destinée aux études acoustico-phonetique et au développement de système de reconnaissance automatique de la parole. (26)

Elle contient un total de 61 phonèmes constituant la phonétique de la langue anglaise tirés de 6300 phrases en anglais-américain, plus exactement 10 phrases prononcées par 630 locuteurs (438 hommes et 192 femmes). Les locuteurs proviennent de 8régions différentes des Etats-Unis. Les enregistrements audio sont encodés en fichiers audio sans compression (wav) et au format 16kHz/16bits.

Le corpus TIMIT comprend des transcriptions orthographiques, phonétiques, et des mots alignés dans le temps ainsi qu'un fichier de forme d'onde de parole 16bits16khz pour chaque énonciation. (27)

3.2.2 Extraction des paramètres acoustique

L'implémentation des HMM nécessite une étape de paramétrage des signaux de parole. Pour celle-ci les MFCC (Mel Frequency Cepstral Coefficients) sont les plus utilisés dans ces domaines, en exploitant les propriétés du système auditif humain par la transformation de l'échelle linéaire des fréquences en échelle Mel.

Les vecteurs MFCC ainsi que leur premier et deuxième coefficients, à savoir MFCC + delta + delta-deltas, sont sélectionnés pour des modèles HMM individuels. Les coefficients delta et delta-delta sont inclus pour rendre le modèle sensible au comportement dynamique du signal. La modélisation acoustique HMM et la segmentation HMM ont lieu avec plusieurs environnements de mélanges gaussiens. La modélisation acoustique avec les mélanges multiples Gaussiens 1,2, 4 et 8 permet d'améliorer considérablement les résultats de la segmentation automatique de la parole, car elle permet d'éviter le problème résultant de l'utilisation du même type de distribution de densité de probabilité pour différents modèles et états. À la fin, les phonèmes avec leurs limites de temps sont obtenus en tant que sortie de segmentation de parole automatique basé sur HMM.

Nous avons manipulé chaque signal en extrayant les MFCC (22):

- Fenêtrage du signal de parole par une fenêtre glissante (Hamming) de durée 25 ms tous les 10 ms ;
- Préaccentuation du signal à l'aide d'un filtre passe-haut ($H\ z$ =1−0.97z−1) ;
- 12 ,14 et 16 MFCC sont extrait de chaque trame de parole ;
- Transformation en échèle Mel à l'aide d'un banc de 26 filtres triangulaires.

3.2.3 Segmentation automatique par HMM

Nous avons implémenté les HMM sous l'environnement Matlab. Le test de ces modèles est fondé sur le corpus choisi. Pour optimiser le modèle, nous avons élaboré et testé plusieurs configurations possibles selon le nombre des composantes gaussiennes (1, 2, 4,8).

Plusieurs étapes sont implémentées pour construire le système ASPh. Ces étapes sont résumées comme suit:

a Préparation des données

La première étape est la construction du fichier de grammaire, pour définir les contraintes de l'entrée du système (28). Il est nécessaire de donner au système des indications pour qu'il puisse déterminer une solution satisfiable. Nous avons construit notre grammaire en suivant le format du HTK. Cette information est stockée dans un fichier texte appelé dictionair2.dic (5)

Le fichier de grammaire est créé avec:

> *$ph = sil dh ix tcl t uw th f er iy f ax gcl g aa tcl t axh kcl k ah m w ix n r aa dcl jh axr z tcl t uw th f eh l aw tcl t sil ;*
>
> *({START_SIL } { $ph } {END_SIL})*

La conversion de fichier de grammaire dans le réseau de treillis de mots HTK est obtenue à l'aide de l'outil «HParse»:

> *$ HParse grammaire wordnet*

Le dictionnaire fournit la correspondance entre les phonèmes utilisés dans le fichier de grammaire et les modèles acoustiques. Dans cette tâche, les modèles phonémiques sont utilisés pour simplifier la structure du dictionnaire qui est représentée comme suit:

> *? [?] ?*
>
> *aa [aa] aa*
>
> *ah [ah] ah*
>
> *aw [aw] aw*
>
> *ax [ax] ax*
>
> *axh [axh] axh*
>
> *axr [axr] axr*
>
> *dcl [dcl] dcl*
>
> *START_SIL [sil] sil*
>
> *END_SIL [sil] sil*

b Extraction de caractéristiques (28)

C'est l'étape d'extraction des paramètres acoustiques MFCC (Coefficients de Cepstral Mel). La méthode MFCC est la meilleure pour la reconnaissance de la parole, et les dérivés primaires et secondaires fournissent des informations supplémentaires.

Ces paramètres sont calculables par le biais d'une fonction dont dispose l'outil HTK. Cette fonction est HCOPY qui prend en entrée un fichier audio et calcule ses coefficients suivant une configuration de la taille des fenêtres, nombre de ceptres, le type de fenêtrage, et d'autres paramètres introduits par l'utilisateur.

Ce paramétrage doit être compatible avec l'outil HTK, la nature des données audio (format, échantillonnage, fréquence, etc.) et les caractéristiques des paramètres (type de paramètre, longueur de la fenêtre, préaccentuation, etc.). Pour ce faire, nous avons créé un fichier de configuration comme suit:

```
# Coding parameters
SOURCEKIND = WAVEFORM
SOURCEFORMAT = WAV
SOURCERATE = 625
TARGETKIND = MFCC_E_D_A
TARGETRATE = 100000
WINDOWSIZE = 250000
USEHAMMING = TRUE
PREEMCOEF = 0.97
NUMCHANS = 26
NUMCEPS = 12
ENORMALISE = TRUE
```

Il est possible de modifier certains paramètres en fonction de nos besoins. Ensuite, un fichier de script HTK "Hcopy.scp" est créé. Il contient les lignes suivantes:

```
../train/S1.WAV ../train/S1.mfc
../train/S2.WAV ../train/S2.mfc
..
```

Chaque fichier du corpus d'apprentissage est représenté par une ligne. Le script 'Hcopy.scp' indique au HTK d'extraire les paramètres acoustiques de chaque fichier audio de la première colonne, puis de sauvegarder les résultats dans la deuxième colonne du fichier correspondant. La commande responsable est:

```
$ HCopy -T 1 -C config -S hcopy.scp
```

c Prototype et initialisation des modèles

Pour l'outil HTK, les chaines de Markov cachées sont d'abord estimées par des prototypes. La fonction d'une définition de prototype est de décrire la forme et la

topologie du HMM, les nombres réels utilisés dans la définition ne sont pas importants. Par conséquent, la taille du vecteur (VecSize) et le type de paramètre (MFCC) devraient être spécifiés et le nombre d'états doit être choisi (NumStates). Les transitions permises entre les états devraient être indiquées en mettant des valeurs différentes de zéro dans les éléments correspondants à la matrice de transition (TransP) et zéros ailleurs. (28)

Un prototype HMM ('proto') est créé, de gauche à droite, avec 3 états et 39 vecteurs de paramètres acoustiques, comme suit:

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "proto"
<BeginHMM>
<NumStates> 5
<State> 2
<Mean> 39
0.0 0.0 0.0 ...
<Variance> 39
1.0 1.0 1.0 ...
.
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Avant de démarrer le processus d'apprentissage, les paramètres des HMMs doivent être correctement initialisés en utilisant la base d'apprentissage afin de permettre une convergence rapide et précise de l'algorithme d'apprentissage.

L'outil HTK proposait le module «Hinit» pour initialiser le prototype avec les moyennes et les variances des données d'apprentissage. Chaque état du HMM comporte plusieurs composantes gaussiennes, les vecteurs d'apprentissage sont associés à la composante gaussienne la plus probable. Le nombre de vecteurs associé à chaque composant dans un état peut être utilisé pour estimer les poids des mélanges.

La commande HInit de l'outil HTK permet d'initialiser les HMMs par alignement temporel en utilisant l'algorithme de Viterbi à partir des prototypes, et les données d'apprentissage dans leur forme MFCC et leur fichier étiqueté associé.

La commande de structure 'Hinit' est donnée par (5):

$ HInit hmm data1 data2 data3

L'organigramme suivant résume le processus:

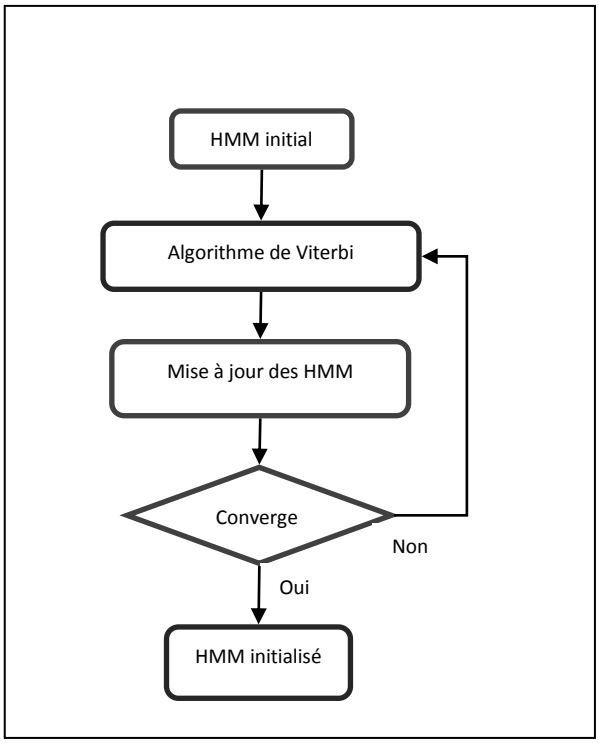


Figure 3.1 : Opération Hinit

Premièrement, HTK charge le prototype du HMM à définir, ensuite il cherche dans la base des étiquettes le label portant le nom de ce HMM ; à noter qu'un fichier label contient le temps de début et de fin d'une étiquette dans un enregistrement. Et par le biais du fichier de configuration il trouve le lien avec les coefficients MFCC calculés précédemment et en prend ensuite ce dont il a besoin. (5)

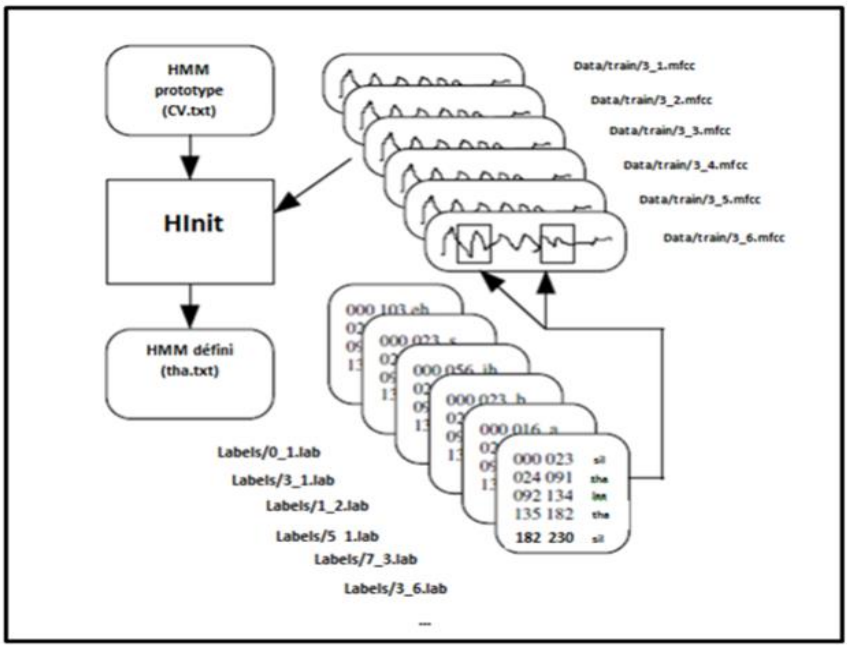


Figure 3.2 : Processus de chargement de données pour la commande Hinit

Quand le système charge tout ce dont a besoin l'algorithme de Viterbi est employé pour trouver l'ordre le plus susceptible d'état correspondant à chaque exemple d'apprentissage, puis les paramètres de HMM sont estimés.

d Apprentissage du système

Et pour l'apprentissage nous allons appliquer l'algorithme de Baum-Welch vu en deuxième chapitre. Le modèle final du HMM est estimé en appliquant la commande «HRest» au modèle généré par «Hinit».

> *HERest -I phones0.mlf -t 250.0 150.0 1000.0 -S train.scp -H hmm0\macros -H hmm0\hmmdefs -M hmm1 mphone0*
> *HERest -I phones0.mlf -t 250.0 150.0 1000.0 -S train.scp -H hmm1\macros -H hmm1\hmmdefs -M hmm2 mphone0*

Cette commande crée un HMM pour chaque phonème donné par le fichier 'phones0.mlf' à partir du modèle stocké dans le dossier 'hmm0'. Les modèles résultants doivent être stockés dans le dossier 'hmm1'. Cette procédure sera répétée trois ou quatre fois. (28)

e Phase de segmentation

Une phase de segmentation est requise après la création du modèle. Ceci est fait en exécutant la commande "Hvite", quand "testf.mfc" est le signal de parole. (28)

> *HVite -H macros -H hmmdefs -i phnfile.phn -I word.mlf -S testf.mfc dict mphone1*

3.2.4 Mesure de performances

La mesure de l'efficacité des deux systèmes proposés pour la SAPh est effectuée en calculant le Taux de segmentation Correct (TSC) en fonction du nombre de composantes gaussiennes. (22)

$$TSC(\%) = \frac{nombre\ des\ trames\ correctement\ segmentées}{nombre\ total\ des\ trames} * 100$$

3.3 Résultats et discussions

Dans cette partie, nous exposons les résultats obtenus dans la segmentation automatique en phonème par HMM. Nous avons implémenté et testé les HMM pour le corpus TIMIT (Anglais).

Trois HMM sont implémentés pour chaque classe 3-HMM, 5-HMM et 7-HMM avec 1, 2, 4 et 8 Composantes Gaussiennes (CG).

3.3.1 Influence des MFCC seuls

Dans un premier temps, nous étudions l'influence des coefficients MFCC sans ses paramètres dynamiques. Nous avons implémenté le système avec 12, 14 et 16 MFCC.

	Nombre de composantes gaussiennes			
Nombre d'états	1	2	4	8
3	61.91%	64.01%	64.20%	64.06%
5	65.77%	64.43%	64.18%	63.61%
7	64.51%	65.06%	64.05%	65.93%

Tableau 3.1 : TSC pour 12 MFCC

Les résultats de ce tableau montrent que pour 12 MFCC, la meilleure performance de 65.93% est obtenue avec un HMM de 7 états et de 8 CG. Cependant, un HMM de 3 états et 1 CG est le moins performant avec un TSC de 61.91%.

	Nombre de composantes gaussiennes			
Nombre d'états	1	2	4	8
3	61.43%	63.39%	63.85%	63.52%
5	57.10%	64.40%	64.07%	62.66%
7	60.19%	65.94%	64.73%	63.28%

Tableau 3.2 : TSC pour 14 MFCC

Ces résultats montrent que pour 14 MFCC, un HMM de 7 états et de 2 CG donne de meilleures performances avec un TSC de 65.94%. En outre, avec un HMM de 5 états et 1 CG les performances dégrade jusqu'à un TSC de 61.91%.

	Nombre de composantes gaussiennes			
Nombre d'états	1	2	4	8
3	60.69%	62.54%	63.66%	63.20%
5	57.45%	63.73%	63.09%	61.28%
7	60.74%	65.16%	65.20%	62.22%

Tableau 3.3 : TSC pour 16 MFCC

Les résultats de ce tableau montre que pour MFCC = 16, la meilleur performance avec TSC=65.20% est obtenue avec un HMM_7 et 4CG et le moins performant, HMM_5 et 1CG avec TSC=61.91%

D'une manière générale, on peut dire que d'après les résultats présentés dans le tableau 2, le meilleur pourcentage obtenu globalement est TSC= 65.94% pour un HMM_7 avec 2 CG ainsi que le plus faible pourcentage TSC = 57.10% pour Hmm_5 avec 1 CG. Ce qui signifie qu'un HMM de 7 états, le système est plus efficace par rapport un HMM de 3 ou de 5 états

3.3.2 Influence des paramètres dynamiques

Les tableaux montrent le taux de classification correcte obtenu en fonction du nombre des Composantes Gaussiennes CG (qui modélise la sortie du HMM), pour des coefficients MFCC incluant les dérivées premières (Δ) et second (ΔΔ).

Nombre des MFCC	**Nombre d'états**	**Nombre de composantes gaussiennes** 1	2	4	8
39	3	70.57%	70.31%	69.38%	67.27%
	5	71.39%	71.40%	67.04%	70.08%
	7	73.91%	70.45%	69.42%	64.51%
45	3	70.34%	69.09%	69.19%	66.66%
	5	71.05%	71.72%	64.78%	69.83%
	7	73.74%	73.95%	68.63%	64.37%
51	3	70.14%	68.68%	69.05%	66.60%
	5	71.81%	70.95%	63.52%	61.51%
	7	73.89%	73.34%	68.61%	64.06%

Tableau 3.4 : TSC en fonction des paramètres dynamiques

Les résultats obtenus dans les tableaux 3-4 montrent que pour les MFCC 12 et 16, on a un TSC maximal avec un HMM de 7 états avec 1 CG, avec des pourcentages respectifs de 73.91% et 73.89%, mais pour MFCC=14 le meilleur TSC obtenu est pour le HHM de 7 états avec 2 CG (TSC = 73.95) représentant ainsi une meilleure performance avec le taux le plus élevé obtenu en général.

Nous constatons globalement, une augmentation dans les performances avec l'implémentation des MFCC en ajoutant les paramètres dynamiques (Δ et ΔΔ) ;

Les expériences décrites dans ce dernier tableau montrent que les informations dynamiques prises en compte dans les modèles HMM modélisant notre système sont pertinentes puisqu'elles permettent d'obtenir des pourcentages nettement plus supérieurs que dans la 1^{ière} partie.

3.3.3 Comparaison de temps d'exécution

On se propose d'évaluer le temps d'exécution du programme dans le but d'évaluer la performance. On note le temps pour la phase de segmentation (data train) ainsi que pour la phase de test (data test). Les tableaux suivants montrent le temps

d'exécution d'un fichier dans la base train et dans la base test en (s) en fonction de nombre de CG.

Nombre des MFCC	**Nombre d'états**	**Nombre de composantes gaussiennes** 1	2	4	8
12	3	100	254	427	782
	5	91	182	318	637
	7	100	200	318	482
14	3	100	182	336	682
	5	91	127	260	364
	7	100	136	216	382
16	3	109	191	236	682
	5	100	136	254	418
	7	109	145	227	391

Tableau 3.5 : Temps d'exécution pour la phase d'apprentissage (s)

Nombre des MFCC	**Nombre d'états**	**Nombre de composantes gaussiennes** 1	2	4	8
12	3	0.51	0.52	0.54	0.48
	5	0.50	0.55	0.59	0.58
	7	0.53	0.58	0.71	0.68
14	3	0.40	0.43	0.47	0.48
	5	0.42	0.44	0.51	0.59
	7	0.49	0.48	0.59	0.75
16	3	0.43	0.46	0.48	0.48
	5	0.44	0.47	0.55	0.62
	7	0.46	0.51	0.59	0.81

Tableau 3.6 : Temps d'exécution pour la phase de segmentation (s)

D'après ces tableaux, on remarque la phase de d'apprentissage implique un temps d'exécution énorme pour que le processus converge. Le temps de segmentation pour chaque phrase est trop faible par rapport à la segmentation manuelle. On Remarque aussi que le temps d'exécution augmente lorsque le nombre de composante gaussienne augmente.

3.4 Conclusion

L'application sous Matlab a permis de mettre en œuvre la méthode HMM pour notre système de segmentation automatique pour cela, nous avons utilisé un programme Matlab, permettant de réaliser les phases importantes de notre système.

Nous avons aussi testé les performances des HMM en calculant pour chaque état de HMM le taux de segments corrects et en évaluant le temps d'exécution du programme pour les 2 phases (phase d'apprentissage et phase de test). Ce qui nous montre que dans plusieurs cas le HHM à états 7 offre la meilleure performance mais avec un temps d'exécution lent.

Le système à base de paramètres dynamiques (MFCC + ΔMFCC + ΔΔMFCC) offre les meilleures performances; rappelons toutefois que ce système opère sur des vecteurs de dimension TARGETKIND = MFCC+1 *3 (39, 45,51) et qu'il s'avère donc plus complexe que le système proposé qui s'appuie sur des vecteurs unitaires. De manière générale la prise en compte de la première et deuxième dérivée des coefficients cepstraux améliore toujours les résultats.

Conclusion générale

Dans ce travail nous avons abordé un domaine en cours d'expansion cette dernière décennie : la segmentation automatique de la parole.

Après avoir lu différents documents sur ce domaine, nous avons pris le choix de travailler avec les chaines de Markov cachées (HMM) qui représentent un outil très robuste en s'appuyant sur des fondements mathématiques très solides et qui se caractérise par la notion d'états/transitions qui permet de traiter les phénomènes temporels dont la parole fait partie. Nous avons opté pour l'outil HTK afin de manipuler les HMMs.

Nous avons réalisé un système de segmentation de la base TIMIT, ce système se compose d'une base d'apprentissage et d'une base de test. Les données sont représentées par des vecteurs caractéristiques de type MFCC. Pour capter certains comportements et évolutions du signal dans le temps, et afin de prendre en compte la dynamique du signal, il nous a semblé nécessaire d'implémenter nos modèles sur un espace de paramètres dynamiques augmenté de la premières et la deuxième dérivée des coefficients cepstraux.

Les tests réalisés nous ont donné les résultats vus au troisième chapitre.

Des recherches plus approfondies sont également nécessaires sur la précision des limites des phonèmes et leur cohérence dans différents contextes de phonèmes et transitions phonétiques entre catégories phonétiques (26).

Il est également nécessaire de recommander aux autres chercheurs de poursuivre les recherches avec une topologie HMM non uniforme pour les modèles acoustiques car la durée des phonèmes est variable. La segmentation automatique de la parole

sans fonction du locuteur est essentielle. Cette segmentation automatique de la parole, indépendante du locuteur, devrait améliorer les performances de la segmentation de parole grâce aux techniques d'adaptation du locuteur (26).

Ce projet nous a permis d'apprendre et surtout de toucher à plusieurs domaines tels que le traitement de signal, la programmation, le traitement de la langue.

Le choix de notre méthode de segmentation (HMM) s'est avéré être une bonne idée et ceux pour les sa capacité à capter les informations nécessaire au traitement et surtout au niveau de l'introduction des dérivées d'ordre 1 et 2 dans les paramètres acoustiques($\Delta_{ck}\Delta\Delta_{ck}$).

Bibliographie

1. **Thomas, Delphine.** carnets2psycho. [En ligne] Proxgroup. [Citation : 15 06 2019.] https://carnets2psycho.net/dico/sens-de-segmentation.html.

2. **Fant, Gunnar.** *Head ,Speech communication and Musical Acoustics.* Stockhilm ,sweden : s.n. pp. 546, 149,257,258,409,414.

3. **Netfi, Samir.** *Segmentation automatique de parole en phones.* Univesité de Rennes. Rennes : s.n., 2006. Thèse de doctorat.

4. **Décodage du Signal de la Parole.** *Overblog.* [En ligne] 2006. http://outilsrecherche.overblog.com/pages/Notes_311_Decodage_du_Signal_de_la_Parole-3082466.html.

5. Ryadh, BENAMMAR. *Traitement Automatique De La Parole Arabe par les HMM :Calculatrice Vocale.* 2012. pp. 18-23.

6. Marie, Tahon. *Traitement de la parole.* 2017. pp. 6,28,29.

7. Sara, ABDELOUAHED. *Analysespectro-temporelle du signal vocal en vue de depistage et du suivie des dyphones chroniques d'origine laryngees.* université Abou Bekr Belkaid Tlemcen. Tlemcen : s.n., 2012. These de Master.

8. HARRAG, Abdelghani. *Extraction des données d'une base: Application à l'extraction des traits du locuteur.* Université Ferhat ABBAS. Setif : s.n., 2011. pp. 59,62, Thése de doctorat.

9. PINQUIER, Julien. *Indexation sonore : Parole / Musique /Bruit.* Universite Paul SBATIER. Toulouse : s.n., 2001.

10. Richard, Celine. *Etude de l'encodage des sons de parole par le tronc cérébral dans le bruit ».* Université de Lyon 2. lyon : s.n., 2010. Thése de doctorat.

11. Abdelhak, SOUIDI. *debruitage de la parole.* Université USTO. Oran : s.n., 2013. cours de master.

12. Bonastre, Jean-François. *In Speaker and Language Recognition Workshop (IEEE Odyssey),.* 2008.

13. RICHARD, Christophe D'ALESSANDRO Gaël. *Synthèse de la parole à partir du texte.* Orsay : s.n., 2013.

14. J paillé, J.P. BAUVIALA, R.CARRE. *Synthése de la parole : description et utilisation d'un synthétiseur du type A FORMANT.* Grenoble : s.n., 1970. pp. 785,786.

15. Poitou, Jacques. *Reconnaissance et synthèse vocales .* 2017.

16. JACQUIER, Caroline. *Étude d'indices acoustiques dans le traitement temporel de la parole chez des adultes normo-lecteurs et des adultes dyslexiques.* 2008. pp. 9,10,11.

17. crouzet. *Segmentation de la parole en mots et régularités phonotactiques : Effets phonologiques, probabilistes ou lexicaux ?* Université Paris 5. paris : s.n., 2000. These de doctorat .

18. Anne Carter, David Cutler. *The predominance of strong initial syllabes in theEnglish vocabulary.* (1987).

19. l'Encyclopaedie, les redacteurs de. Britannica. *Consonnant Phonetics.* [En ligne] https://www.britannica.com/topic/consonant.

20. Chauhan, Yamini. *ÉCRIT PAR Les rédacteurs de l'Encyclopaedia Britannica.* 2013.

21. Lety, Monique. *Transcriptionorthographique-phonétique:unsystème interpréteur.* Université de Grenoble. Grenoble : s.n., 2006. Thèse de doctorat.

22. Abed Ahcene, Amrouche Aissa, Delmadji Abdelkader, Boubakeur Khadidja, Droua-Hamdani Ghania. *Segmentation Automatique des Signaux Sonores par HMM et RNA pour la langue Arabe.* 2016. pp. 1,2,3.

23. Scheffer, Nicolas. *Techniques et applications de traitement de la parole.* 2003.

24. AMARA, Fatima Zahraa , SERIARI, Meriem. *La reconnaissance des battements cardiques par les HMMs.* Universite Abou Bekr Belkaid Tlemcen UABT. tlemcen : s.n. pp. 24,27,30-33, Diplome de Master 2.

25. Matthieu Petit, Christiansen Henning. *Un calcul de Viterbi pour un Modèle de Markov Caché Contraint.* **2009. pp. 3,4.**

26. Lamia, AZIB. *Apllication des Modèles de Markov Cachés et les Modéles de Melanges de Gaussiennes par classification phonetique.* UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE D'ORAN Mohamed Boudiaf. Oran : s.n., 2012. p. 51, diplôme de Magister.

27. Lamare, François. *Segmentation non supervisée d'un flux de parole en syllabes.* Institut de Recherche et Coordination Acoustique/Musique (IRCAM). Paris : s.n., 2012. p. 29, Master 2 recherche .

28. Abed Ahcène, Amrouche Aissa and Boubakeur Khadidja Nesrine. *Investigation of HTK for Arabic Phonemes Boundary Detection.* 2017. pp. 272-276.