الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم و البحث العلمي
MINISTERE DE L'ENSEIGNEMENT ET DE LA RECHERCHE SCIENTIFIQUE

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT : d'Electronique

PROJET DE FIN D'ETUDES

SUJET

La segmentation en Traitement Automatiqne de la Parole

Proposé Par : Melle M. GUERTI

Etudié par : B. GUELLOUR

Dirigé par : Melle M. GUERTI

PROMOTION : Juin 1988

E.N.P. 10, Avenue Hacen Badi – EL-HARRACH – ALGER

****** D E D I C A C E S *******

A mes grands parents

A mes parents

A toute ma famille

A tous(toutes) mes amis (es)

BOUALEM

****** R E M E R C I E M E N T S ******

Je tiens à remercier vivement mon promoteur Melle M. GUERTI pour son aide et ses conseils qu'elle m'a prodigués tout au long de la réalisation de ce travail.

Je suis trés reconnaissant à Mrs A. BOUAFIA, B. TOUATI, et J. DAUCET pour leur aide.

Que tous les professeurs, qui ont contribué à notre formation, trouvent ici l'expression de ma profonde gratitude.

***** S O M M A I R E *****

INTRODUCTOIN GENERALE

Les machines qui parlent vont avoir beaucoup d'applications utiles dans le futur. Beaucoup de temps pourrait-être économisé, en utilisant un ensemble d'instructions parlées à la place d'un ensemble d'instructions imprimées, quand il faudra mettre en marche un équipement complexe; dans les situations où les yeux et les mains sont déjà occupés, où des actions peuvent-être representées sous forme d'informations nouvelles qui peuvent être mieux représentées sous forme de messages parlés.

La première "machine à parler" a été réalisée au 18(ème) siècle par WOLFGANG Von Kemplen; il savait lui faire parler plusieurs mots clairement et distinctement.

De nombreux systèmes avec une sortie d'informations sous forme de parole, dépendent de l'utilisation de messages complets enregistrés à l'avance. Le système serait plus flexible et offrirait de plus larges perspectives si les messages parlés pouvaient-être construits à partir d'unités plus petites de la même façon que le langage écrit. Ces unités constituent les éléments de base qui pourrait-être un alphabet du langage parlé par analogie avec l'alphabet du langage écrit. Un système avec un tel ensemble d'éléments de base pourrait donner une possibilité de vocabulaire illimité à porté de la main. La segmentation, consiste en la division du signal de parole en segments de telle manière que chaque

segment porte en lui même une signification. L'intérêt de la segmentation est l'obtention de segments de parole bien délimités. Ces segments sont les unités de base du traitement automatique de la parole.

Le but de notre étude est le choix et la construction de cet ensemble d'éléments de base et la présentation d'un modèle pour prévoir les durées des phonèmes dans le langage. Pour atteindre ces objectifs, le travail a été mené de la manière suivante:

Le premier chapitre, traite les phénomènes de production de la parole et les différentes unités de base du langage parlé.

Le second est consacré à l'étude du schéma général de la communication parlée, ainsi que les problèmes relatifs au traitement automatique de la parole.

Au troisième, nous présentons quelques méthodes de segmentation automatique de la parole.

Dans le dernier chapitre, nous présentons un modéle de durées segmentales pour une application en synthèse pour le français.

En annexe nous présentons le programme de prédiction des durées segmentales écrit en Basic sur Rainbow 100 de l'E.N.P.A.

CHAPITRE I

LA PAROLE NATURELLE

I-1 INTRODUCTION

Le phénomène de la production naturelle de la parole suscite aujourd'hui l'intérêt des chercheurs. En effet, une meilleure connaissance de ce phénomène est devenue nécessaire pour pouvoir avancer dans les secteurs de la synthèse et de la reconnaissance de la parole. Une telle connaissance exige l'étude de plus en plus approfondie des caractéristiques acoustiques de la parole; ainsi que celle des différentes divisions possibles du signal de parole.

I-2 PRODUCTION DE LA PAROLE NATURELLE

Le discours de la parole est une succession de différences de pression de l'air qui sont engendrées par le système de production de la parole humaine (fig I-1).

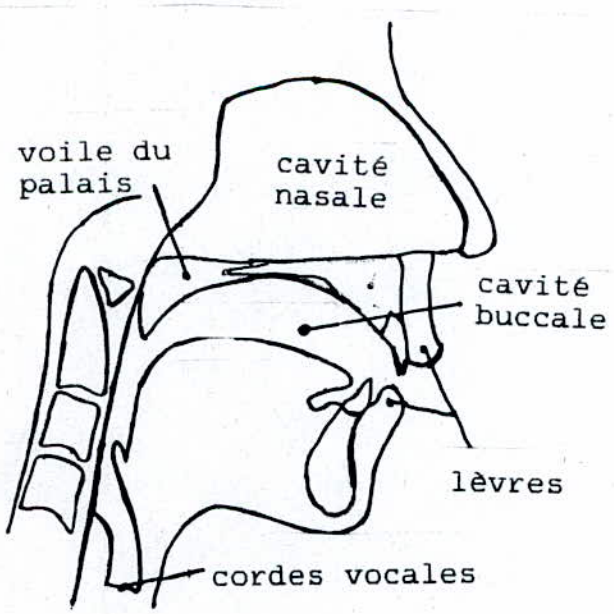


Figure I-1 Système phonatoire humain

La bouche, le pharynx et la cavité nasale (le conduit vocal) peuvent-être considérés comme une cavité résonnante qui est excitée par les vibrations périodiques des cordes vocales, et transformée dans les éléments vocaux du langage (par exemple la voyelle /a/), ou par des turbulances qui prennent naissance dans des rétrécissements ou des constrictions du conduit vocal (par exemple la consonne /s/).

I-3 CARACTERISTIQUES ACOUSTIQUES DU SIGNAL DE PAROLE

L'excitation qui produit la parole peut-être caractérisée par trois paramètres:

-L'amplitude: qui est en relation avec l'intensité du signal

-Un paramètre qui indique si le son est voisé ou non.

-La fréquence fondamentale (Pitch): la fréquence à laquelle les cordes vocales vibrent.

La cavité résonnante peut-être caractérisée par les fréquences centrales et les largeurs de bande des formants, qui sont les pics de résonance dans la fonction de transfert du conduit vocal (fig I-2). La fréquence fondamentale (Fo) est trouvée dans la structure fine du spectre, les formants dans l'enveloppe. Le spectre de la parole jusqu'à une fréquence de 5 KHz contient habituellement 5 formants, de façon que la cavité résonnante puisse-être décrite par dix paramètres (les cinq fréquences centrales et les largeurs de bande correspondante).

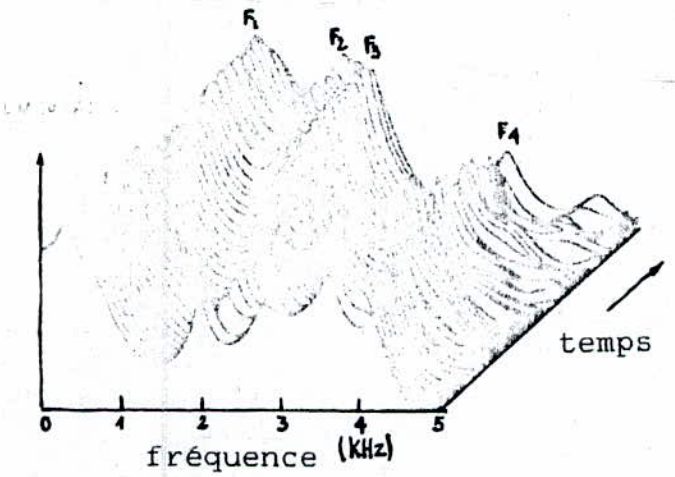


Figure I-2 Spectre vocalique présentant quatre formants

I-4 LES UNITES DE BASE POUR LA SYNTHESE

En principe, la parole synthétique peut-être engendrée en spécifiant le comportement des 13 paramètres (ch I-3), comme une fonction du temps, mais ceci requiert un trés grand nombre de règles compliquées dont certaines d'entre elles sont encore inconnues. En pratique cette approche est difficile et la qualité de la parole est trés mauvaise.

Une autre façon possible pour engendrer la parole synthétique consiste a concaténer (c-à-d joindre les unes à côté des autres) des fragements de parole naturelle. Dans sa forme la plus simple ceci revient à la présentation d'un message complet, audible, enregistré à l'avance. Mais c'est beaucoup plus intéressant de concaténer des

mots séparés pour former des phrases complètes; mais si on envisage toute les possibilités qu'on peut concevoir avec un tel système une liste énorme de mots serait nécessaire .

On peut donc limiter le volume de ce nombre d'éléments de base à enregistrer dans une certaine mesure en utilisant des syllabes que des mots comme unités de base. Mais même ainsi on aura besoin d'une bibliothèque assez importante d'unités de base.

La bibliothèque peut ainsi être réduite à son minimum, si on prend comme unité de base le plus petit segment de parole qui porte en lui même les différentes significations (le phonème). Cependant pour maintenir le caractère courant de la parole, il est nécessaire d'engendrer la transition d'un phonème au phonème suivant par un ensemble de règles qui sont assez difficiles à établir. Nous n'avons pas besoin de telles règles si nous prenons les diphones comme unités de base. Le nombre d'unités de base est maintenant plus grand, mais nous n'ajoutons aucune règle pour les transitions lors de la synthèse (fig I-3).

I-5 LES UNITES PHONETIQUES

Les linguistes ont defini le phonème comme étant la plus petite unité phonètique. La langue française contient 37 phonèmes, comprenant 16 voyelles, 18 consonnes, et 3 semi voyelles (fig I-4). Les critères de classification de ces éléments sont les suivants:

-L'opposition sourde-sonore.

-L'opposition orale-nasale.

-Le lieu d'articulation: qui correspond à l'endroit de la constriction maximale du conduit vocal pour les consonnes. Pour les sons vocaliques, ce lieu correspond plutôt à la position du corps de la langue.

-Le mode d'articulation: qui représente le mécanisme de formation du son.

Le diphonème désigne l'ensemble formé par l'association de deux phonèmes; il est également appelé diphone, phonatome, ou élément phonétique. Le nombre de diphones est pour le français égal au nombre d'arrangements des 37 phonèmes pris deux à deux soit 1332 diphones.

Tout message est succeptible d'être découpé en diphones. A l'inverse, pour reconstituer un mot ou une phrase, il suffit d'assembler les diphonèmes nécessaires par leur extrémité commune (fig I-5).

Il faut noter, que si un mot ou une phrase contient (n) phonèmes, le nombre de diphones correspondant est (n+1) (fig I-5).

Le mot	PARIS
Les phonèmes	/p/ /a/ /R/ /i/
Les diphones	/#p/ /pa/ /aR/ /Ri/ /i#/

Figure I-5 Les phonèmes et les diphones du mot "PARIS"

UNITE DE BASE	AVANTAGES	INCONVENIENTS
MOT	coarticulation incluse indépendant de la langue	grands vocabulaires adaptation au locuteur
SYLLABE	facile à localiser (voyelle) coarticulation incluse	nombre total élevé frontières difficiles à localiser
DIPHONE	contient une partie de coarticulation aucune règle lors de la synthèse	nombre total (1000) problèmes de segmentation
PHONEME	nombre total peu élevé codage aisé des mots dans le lexique	trés dépendants du contexte pas facile à localiser algorithmes et règles complexes pour segmentation et synthèse

Figure I-3 Unités de base de la synthèse

	plosive		fricative		liquide	nasale	semi-voy
	sonore	sourde	sonore	sourde	sonore	sonore	sonore
vélaire	g	k					
dental	d	t	ʒ	s	l	n	
bilabial	b	p				m	w/y
labio-dental			v	f			
palato-alveolaire			z	ʃ			
uvélaire					R		
palatale							j

Figure I-4 Les consonnes de la langue française

I-6 CONCLUSION

L'étude de ces notions est la base de toute étude portant sur le traitement automatique de la parole. En effet, l'apport des experts phonéticiens est évidemment capital.

CHAPITRE II

LA COMMUNICATION PARLEE HOMME MACHINE

II-1 INTRODUCTION

Il convient de séparer les systèmes de synthèse de parole, qui sont des appareils qui parlent, des systèmes de reconnaissance, capables de reconnaître et de comprendre ce qui a été dit, ou de reconnaître la personne qui a parlé.

Nous étudierons dans ce chapitre le schéma géneral de la communication parlée homme-machine, ainsi que les problèmes relatifs au traitement automatique de la parole.

II-2 HISTORIQUE

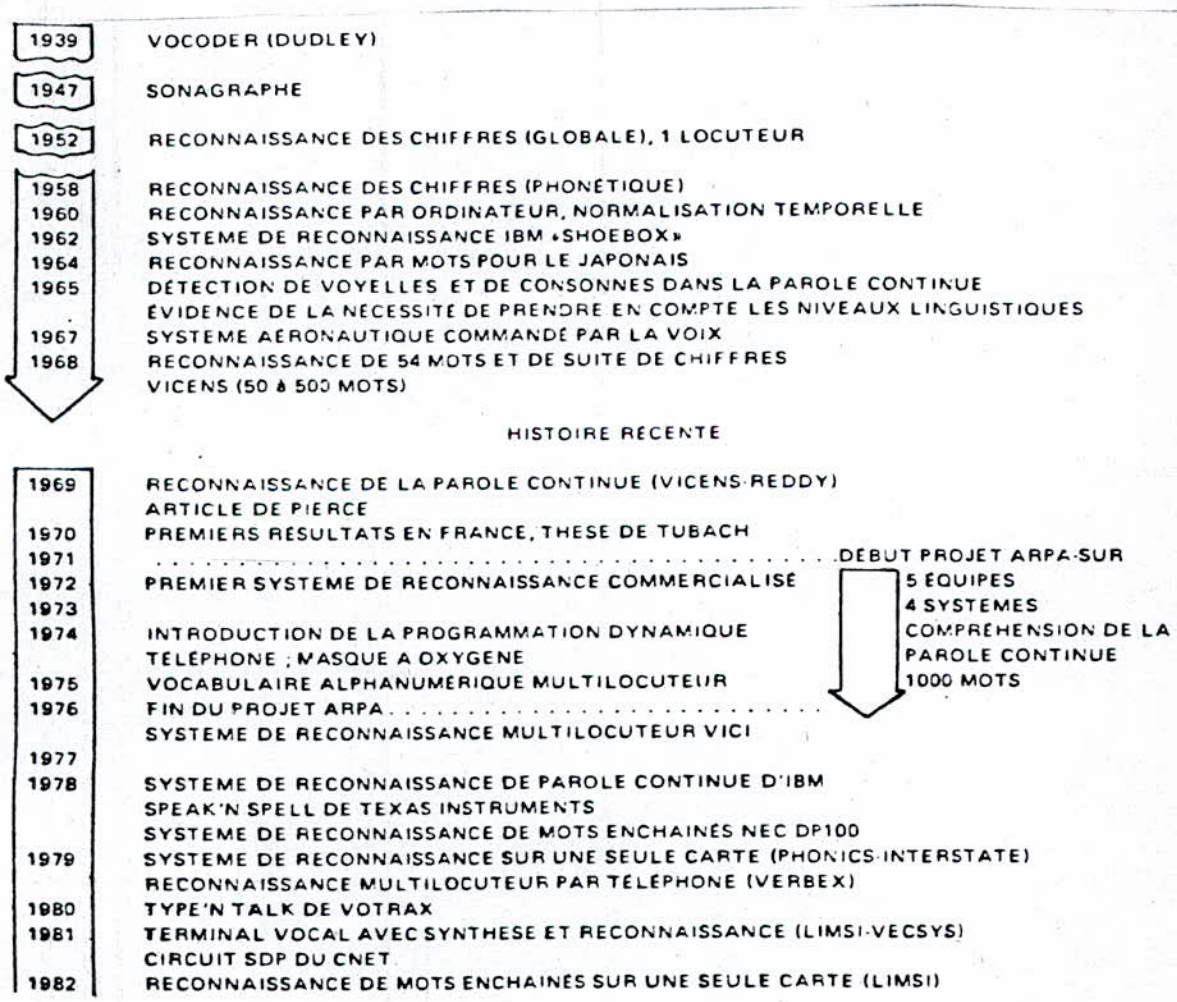


Figure II-1 Historique du traitement automatique de la parole

On peut faire remonter les travaux en matière de traitement automatique de la parole à 1939 (fig II-1), date à laquelle DUDLEY, chercheur aux laboratoires BELL (USA), présenta son VOCODER (ou Voice Coder) capable de coder la parole puis de la restituer. Sur la même base, il présenta ensuite le VODER, premier système de synthèse de parole électrique actionné par un clavier.

En 1947, la réalisation du sonographe permit une visualisation lisible du signal vocal dans ses coordonnées temps-amplitude-fréquence (sonogramme fig II-2), et entraîna de nombreuses études du matériel vocal, en particulier dans les laboratoires de phonétique.

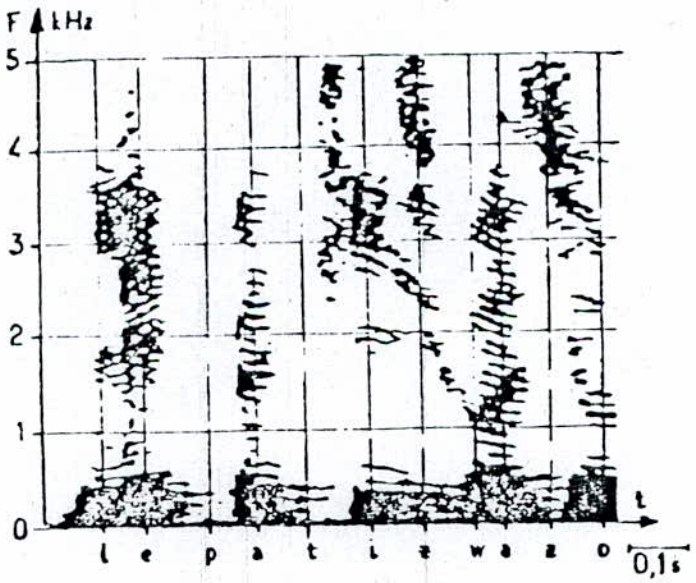


Figure II-2 Sonogramme de la phrase
"les petits oiseaux"

-l'amplitude est marquée par le degré de noirceur
-les traits foncés correspondent aux harmoniques
-les zones les plus sombres aux formants.

Les premiers travaux de reconnaissance apparurent dans les années 50 en Suisse; ces travaux utilisaient des moyens d'électronique analogique, l'utilisation des ordinateurs apparaissant au début des années 60. Le premier système de reconnaissance commercialisé en 1972 par la société THRESHOLD aux USA, consiste à reconnaitre mot par mot des mots isolés prononcés par une seule personne. L'avènement des microprocesseurs a conduit au début des années 80 à une miniaturisation de ces systèmes et à une diminution considérable de leur prix.

II-3 SCHEMA GENERAL DE LA COMMUNICATION PARLEE HOMME-MACHINE

Dans le cadre de la communication parlée homme-machine, on peut définir les organes de reconnaissance et de synthèse reliés à un module chargé d'interpréter le sens des messages émis par l'utilisateur humain en un langage de commande compréhensible par l'effecteur qui agit sur la tâche, ou réciproquement de traduire les contrôles qui viennent de la tâche en une phrase bien prononcée.

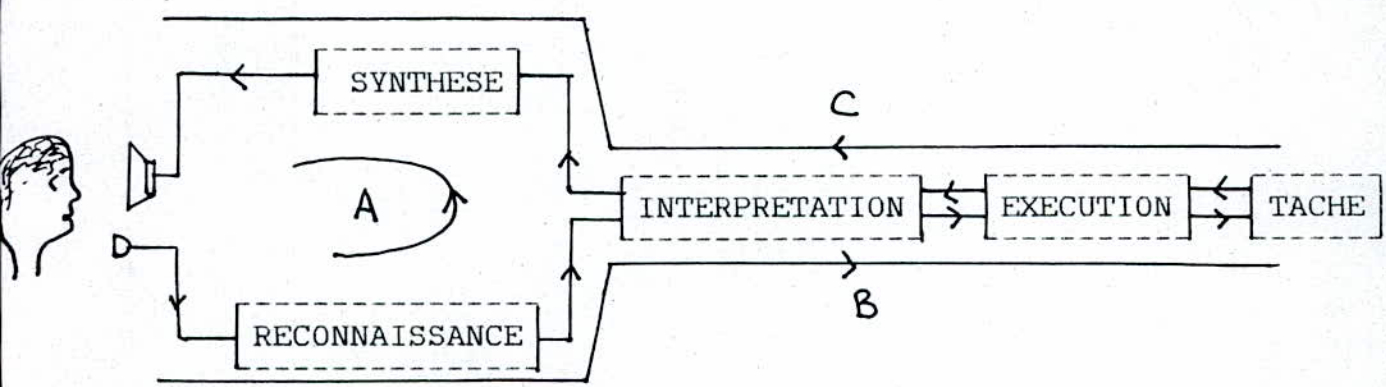


Figure II-3 Schéma de la communication parlée homme-machine

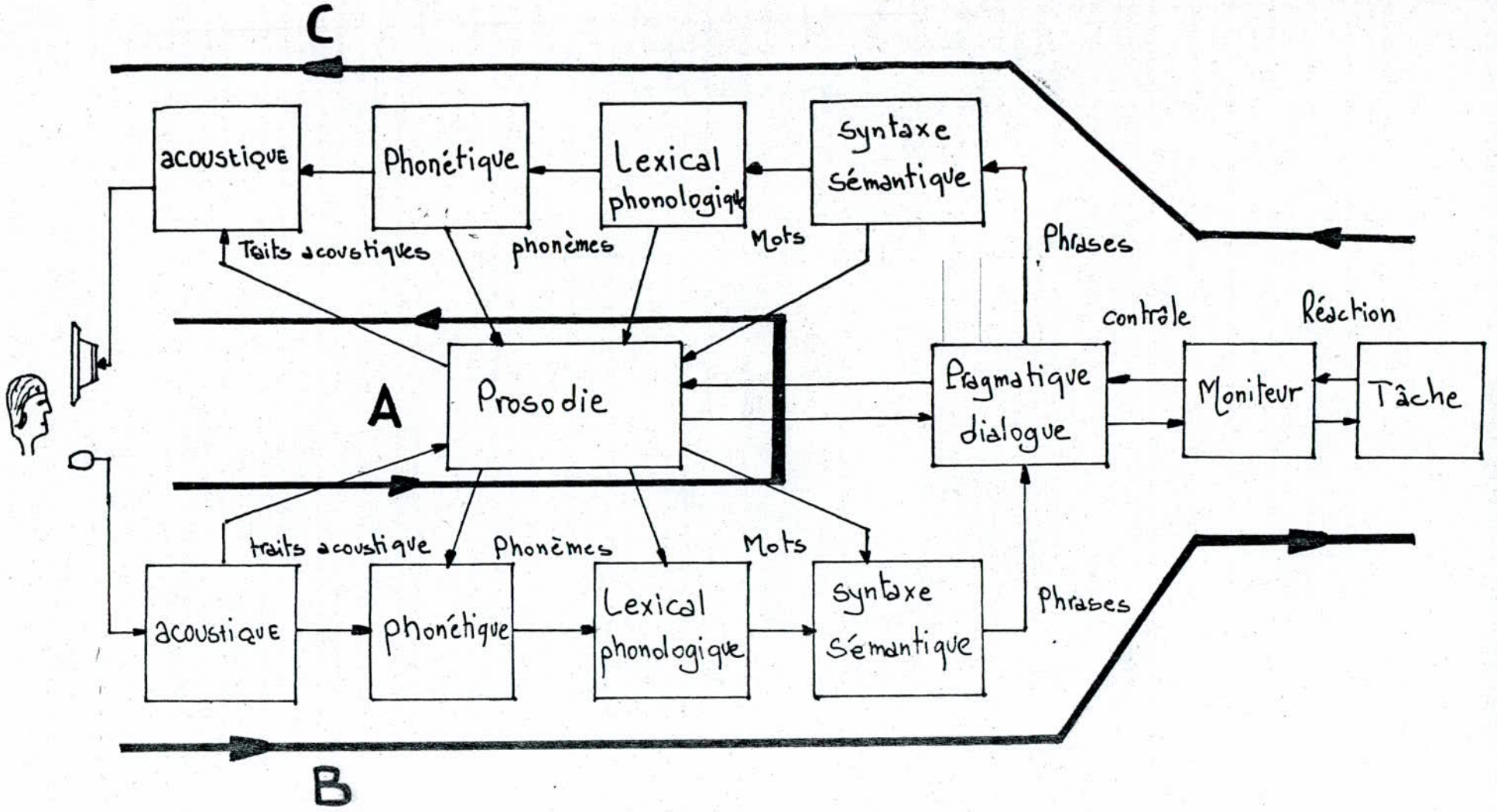


Figure II-4 Schéma géneral de la communication parlée homme-machine

On distingue ainsi plusieurs circuits de communication (fig II-3).

-Le circuit A : ou circuit de conversation, sans action de la tâche (c'est le cas de la prononciation par l'utilisateur d'une phrase dont le sens est ambigu et pour laquelle le système demande une précision).

-Le circuit B : ou circuit de commande.

-Le circuit C : ou circuit de contrôle.

Si l'on analyse plus finement les parties reconnaissance et synthèse on voit apparaître les sources de connaissance qui participent à l'encodage et au décodage de la parole et qui sont les mêmes pour ces deux opérations (fig II-4).

La communication parlée intervient en parallèle avec d'autre canaux de communication (visuel,etc...).

II-4 LES SOURCES DE CONNAISSANCE

II-4-1 Niveau acoustique

Aprés transformation de l'émission vocale, qui est une variation de pression de l'air en signal électrique par un microphone, le niveau acoustique va traduire ce signal analogique en un ensemble d'informations, qui vont permettre la reconnaissance du signal par l'ordinateur. Quatre opérations sont effectuées à ce niveau.

a)Paramétrisation: Le signal analogique est transformé en signal numérique afin de pouvoir être traité par la machine digitale. Cette opération est effectuée par un convertisseur analogique-digital.Le signal numérique est alors analysé par différentes méthodes pour pouvoir être exploité: méthodes dites temporelles (passage par zéro,etc...) ou fréquentielles (transformée de Fourier, banc de filtres numeriques, méthodes de prédiction linéaire,etc...).

Dans le premier cas, le signal est codé dans les coordonnées temps-amplitude, dans le second il est codé dans les coordonées amplitude-temps-fréquence et délivre un vecteur à intervalles réguliers. Cette opération permet de réduire considérablement la quantité d'information transmise en conservant l'information essentielle et en éliminant la redondance inhérente au signal de parole.

b)Segmentation: Segmenter le continuum du signal vocal en les différents sons qui le constituent.Nous revenons à ceci dans les chapitres qui suivent.

c)Extraction des paramètres pertinents: Retenir que les informations qui serviront a l'application: spectre aux instants de plus grande stabilité, ou aux instants de transition par exemple.

d)Informations relatives à la prosodie: Ces informations sont

de trois types: variations du fondamental de la voix (correspondant à la variation de la fréquence de vibration des cordes vocales), intensité (amplitude des sons émis) et rythme (durée des syllabes).Elles sont transmises au niveau prosodique .

II-4-2 Niveau phonétique

Le but est de traduire les suites de traits pertinents en une suite de phonèmes. Pour le français 37 phonèmes existent (cf.ch I-5).

Les problèmes de coarticulation font que la reconnaissance de chaque phonème fait intervenir la reconnaissance des phonèmes contextuels, et des transitions permettant de passer de l'un à l'autre.

II-4-3 Niveau phonologique

Les phénomènes de la langue qui font que le contenu phonétique des mots sera modifié dans une articulation rapide, ou par une succession de termes lexicaux. Ces règles souvent induites par les difficultés de prononciation de certaines suites de phonèmes. On trouvera ici les règles de liaison:

-Liaison interdite: le petit ‿ est malade.

-Liaison obligatoire: le petit ‿ homme est malade.

-Liaison facultative: je suis ‿ arrivé à une heure.

-Les élisions comme celle de E caduque (petit prononcé

p'tit), la dénasalisation (bon ami prononcé bonne ami)etc..

Ce niveau traite également des variétés dialectales que l'on pourra trouver en fonction de l'appartenance géographique ou socioéconomique du locuteur.

II-4-4 Niveau lexical

A ce niveau sont contenues les informations sur les mots qui composent la langue. Un dictionnaire comme le Petit Larousse, comporte environ 20.000 mots communs.La conjugaison des verbes aux différents modes, temps et personnes portent ce nombre à environ 200.000 éléments. A ces éléments il faut ajouter les noms propres, les mots rares tels ceux des vocabulaires techniques. Chacun de ces mots a une ou plusieurs prononciations, en fonction des règles phonologiques, qui sont représentées par leur écriture phonétique. Le problème est que certains mots s'écrivent de la même façon mais se prononcent différemment (homographes hétérophones: "des portions" et nous "portions"), et réciproquement (homophones hétérographes: "une portion" et "des portions"). Chaque mot doit être accompagné de sa catégorie grammaticale, de son genre, de son nombre, du temps pour les verbes etc.....

II-4-5 Niveau synthaxique

Pour le langage naturel,ce niveau est supposé renfermer les règles de la grammaire qui permettent de décrire et

d'analyser le langage, en termes d'analyse grammaticale et fonctionnelle. Outre que la formalisation des règles syntaxiques du langage écrit est loin d'être acquise, il apparaît que les règles d'articulation qui régissent la langue parlée sont encore beaucoup plus floues. De plus, il apparaît qu'analyse syntaxique et analyse sémantique (liée au sens de la phrase) sont fortement liées (les phrases "la fermière vend sa vache parce qu'elle n'a plus de lait" et "la fermière vend sa vache parce qu'elle n'a plus d'argent").

On se contentera donc, dans l'utilisation des technologies vocales, d'utiliser des structures syntaxiques propres à l'application visée. Par contre, des grammaires probabilistes de la langue naturelle existent déjà, qui peuvent être utilisées par des systèmes de machines à écrire à entrée vocale, où il s'agit de traiter la langue écrite.

II-4-6 Niveau sémantique

Ce niveau traite du, ou des sens, des mots tels que l'on peut les trouver dans un dictionnaire, et les relations entre eux. C'est ainsi qu'un être humain pourra comprendre la phrase:
"j'ai voulu faire un petit galop, mais l'écurie était vide" sans que le mot "cheval" ait été prononcé. Des tentatives d'élaboration de tels réseaux sémantiques ont été faites, sans

qu'aucune réalisation puisse être utilisée à l'heure actuelle.

II-4-7 Niveau pragmatique et dialogue

Ce niveau a pour tâche de déterminer le sens d'une phrase, donc de la comprendre, dans le contexte de l'application. Ainsi il déterminera le sens du mot "tableau" suivant que l'action se passe dans une salle de cours, ou dans une galerie de peinture. De même dans le contexte d'une conversation, il lui faudra mettre en rapport les termes décrivant un même élément (tels que "tableau" et "le" dans le dialogue:

-Que veux-tu faire de ce tableau ?

-Je veux le vendre) .

II-5 LES PROBLEMES RELATIFS AU TRAITEMENT AUTOMATIQUE DE LA PAROLE

Les problèmes qui rendent difficile le traitement automatique de la parole sont les suivants:

-Chaque son élémentaire (ou phonème) est déformé par les sons qui le suivent ou qui le précèdent (effet de coarticulation).

-IL n'y a pas de silence entre les mots (comme les blancs qui séparent les mots en langue écrite).

-Une trés grande variabilité est présente dans la parole:variabilité propre au locuteur, comme sa façon de parler (en chantant, en

criant, en chuchotant,etc...), ou variabilité entre locuteurs (voix d'homme, de femme, d'enfant). Les problèmes de prise de son (différents microphones) ou d'environnement (bruit) rajoutent une difficulté supplémentaire.

-Le même signal véhicule plusieurs types d'informations (sur les sons prononcés, sur la structure grammaticale de la phrase, sur l'identité de la personne qui parle, sur son état d'émotivité, ou de santé,etc....).

-Il n'y a pas, à l'heure actuelle, de règles précises permettant de formaliser les connaissances qui interviennent pour décoder le signal de parole (comme la grammaire de la langue écrite; par exemple), et de plus les différentes informations qui participent à ce décodage sont étroitement liées (par exemple le sens d'un mot et sa catégorie grammaticale).

-Il est nécessaire d'observer une trés large quantité de données pour comprendre où se trouve les invariants (ce qui fait que le son /a/ est toujours reconnu comme tel, quel que soit le locuteur, sa façon de parler, le mot dans lequel il se trouve, les conditions de prise de son,etc....).

-Enfin, la réalisation de systèmes d'entrée/sortie vocale demande des connaissances dans des disciplines trés diverses (physiologie, traitement de signal, linguistique, phonétique, intelligence artificielle,etc...).

II-6 CONCLUSION

Les technologies vocales sont donc un domaine en plein essor. Certains aspects sont relativement bien definis (par exemple: les méthodes d'analyse) . Néanmoins le champs des recherches est vaste et les années à venir devraient voir l'éclosion progressive de systèmes toujours plus performants.

CHAPITRE III

SEGMENTATION AUTOMATIQUE DE LA PAROLE

III-1 INTRODUCTION

La segmentation est un des problèmes les plus difficiles à résoudre. La situation idéale serait celle ou chaque segment corresponderait à un phonème. Différentes méthodes existent basées sur les courbes de variation d'énergie, ou de variabilité du signal.

Dans ce chapitre, nous nous sommes particulièrement intéressés à la localisation des frontières des diphones à partir des frontières des phonèmes.

III-2 SEGMENTATION DE LA PAROLE EN DIPHONES

Une bibliothèque de diphones peut-être obtenue de la façon suivante:

Des enregistrements sur bande magnétique sont faits sur un grand nombre de mots qui sont prononcés par un locuteur sélectionné. La liste des mots est choisie de telle manière qu'elle contienne toute les combinaisons possibles des sons de la parole,des mots sans aucune signification peuvent-être utilisés. La procédure est la suivante:

Les mots sont numérisés à un taux d'échantillonage de 10 KHz, ensuite les paramètres mentionnés plutôt (ch I-2) sont détérminés et enregistrés comme une série continue dans une mémoire; ceux-ci sont utilisés pour engendrer (resynthétiser) un autre signal. Les

frontières des phonèmes sont déterminées à partir des valeurs des paramètres et en écoutant le signal de synthèse. Ces frontières ou ces limites sont donc établies d'une manière non automatique.

Les limites des diphones sont calculées à partir des limites des phonèmes en utilisant un ensemble de règles. Les règles pour positionner les limites des diphones sont différentes pour chaque phonème (ch III-4).

Cette méthode non automatique pour construire une bibliothèque de diphones est laborieuse et prend beaucoup de temps. Des bibliothèques de diphones pour plusieurs langues ont été préparées de cette manière pour un locuteur. La qualité de la parole de synthèse produite par la concaténation des diphones à partir de ces bibliothèques est plutôt bonne. Il y'a un grand besoin de bibliothèques de diphones pour plusieurs langues et plusieurs locuteurs pour chaque langue (bien qu'il ait des différences considérables entre les bibliothèques de diphones pour des langues différentes, le principe de concaténation des diphones peut-être utilisé pour des langues différentes), la procédure de préparation des bibliothèques de diphones décrite ci-dessus doit être repétée pour chaque langue et pour chaque locuteur. Ceci est la raison pour laquelle des efforts ont été faits pour trouver une méthode automatique pour la préparation des bibliothéques de diphones à partir d'enregistrement de mots. A coté de l'argument temps, il y'a une seconde raison pour rechercher une segmentation automatique. Une

méthode automatique est supposée donner un résultat final qui sera plus facile à reproduire et plus cohérant qu'une méthode non automatique.

III-3 SEGMENTATION AUTOMATIQUE DE LA PAROLE EN PHONEMES

Les méthodes de segmentation en phonèmes décrites dans la littérature peuvent-être classées en deux groupes:

-D'une part il y'a les méthodes qui cherchent des caractéristiques reconnaissables dans le signal de la parole; comme les frontières entre les états plus ou moins stables dans le signal. Dans les méthodes de cette catégorie, aucunne reconnaissance n'est faite; du fait que le mot à segmenter est déjà connu, il n'y a donc pas d'identification phonétique (c-à-d aucunne tentative de relier un phonème avec un segment associé du signal).

-D'autre part il y'a des méthodes qui font la division sur la base de correspondance entre une partie du signal de parole et une référence de forme connue pour chaque partie du signal.

Chaque méthode a ses avantages et ses inconvenients. Quand une méthode de la première catégorie est utilisée, les limites entre les segments sont définis d'une manière précise, mais les segments délimités par deux frontières ne sont pas donnés avec la moindre identification phonétique. Une méthode de la seconde catégorie ne pourrait satisfaire les nécessités d'un système de synthèse pour ce qui est de la précision des limites ou des frontières.

III-3-1 Segmentation en segments d'état stable

En général, les comportements des propriétés acoustiques du signal de parole comme une fonction du temps sont assez imprévisibles. Cependant il est possible d'identifier des segments d'état stable du signal de parole qui sont habituellement décrits en terme de changement de pression de l'air comme une fonction du temps. Par exemple au point A et B du signal du mot "nenaanene" (mot sans signification) (fig III-1), des transitions peuvent-être vues entre deux segments d'état stable. Ces transitions sont aussi visibles dans la composition spectrale du signal aux différents temps (fig III-2).

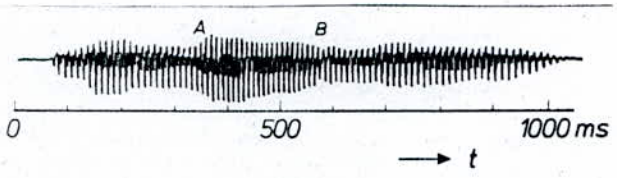


Figure III-1 Signal du mot "nanaanene"

La composition spectrale montre un nombre 'd'images' de spectres d'amplitude à un intervalle de 10 ms. Les formants, les pics dans le spectre, sont sujets à des changements. La fréquence centrale et la largeur de bande changent avec le temps. Avec l'aide d'une telle représentation, les spectres d'amplitude à un intervalle de 10 ms, le signal de parole peut-être divisé en segments d'état stable.

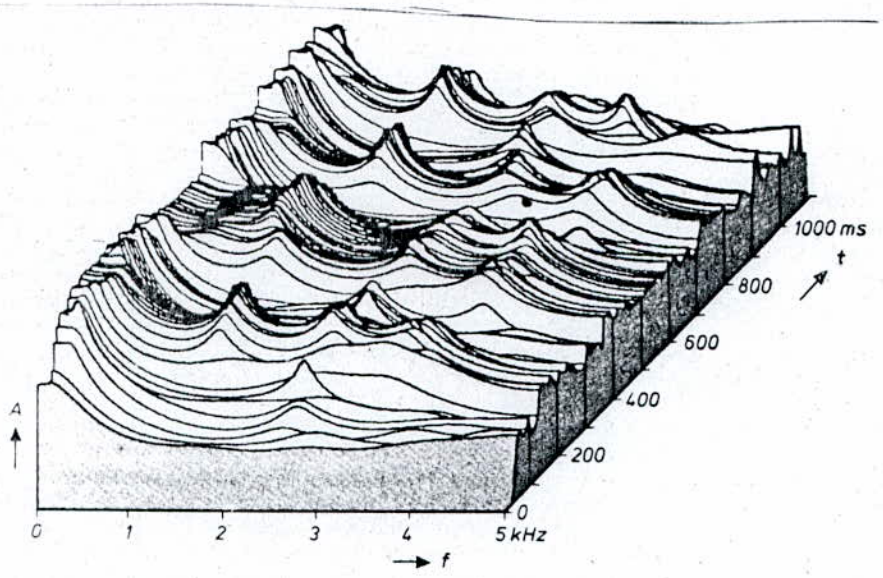


Figure III-2 Spectre du mot "nenaanene"

Deux spectres appartiennent au même segment d'état stable quand ils se ressemblent l'un à l'autre dans une certaine mesure. Nous devons alors indiquer des mesures de similarité entre deux spectres. Pour faire ceci nous allons procéder comme suit:

On détemine la mesure de similarité entre deux spectres en calculant la corrélation entre eux. Cette corrélation Ci,j est définie par:

$$Ci,j = \frac{Si \,.\, Sj}{\left[(Si.Si) \,.\, (Sj.Sj)\right]^{1/2}} \qquad (1)$$

$$\text{où} \quad Si.Sj = \int_{0}^{5KHz} W(f) \cdot Si(f) \cdot Sj(f) \; df \qquad (2)$$

est le produit scalaire des spectres Si et Sj ; f est la fréquence et W(f) la fonction de pondération spectrale qui approxime la sensibilité de l'oreille humaine, cette fonction de pondération spectrale est donnée par:

$$W(f) = \begin{cases} 0 & \text{si} \quad 0 \text{ Hz} < f \leqslant 200 \text{ Hz} \\ 1 & \text{si} \quad 200 \text{ Hz} < f \leqslant 1000 \text{ Hz} \\ 1000/f & \text{si} \quad f > 1000 \text{ Hz} \end{cases} \qquad (3)$$

Si deux spectres sont identiques, la corrélation Ci,j est égale à 1. Mais comme la similarité des spectres diminue, Ci,j va différer de 1. La corrélation entre le i(éme) spectre et ses 10 voisins de chaque côté ($i-10 \leqslant j \leqslant i+10$) est montrée schématiquement dans la figure III-3. La courbe a un maximum 1 pour $i = j$. Aprés avoir choisi une valeur de seuil Ct, on peut identifier un segment qui contient le spectre (i), dans la mesure où la série des spectres succesifs (de ib à ie) dont la corrélation avec le spectre (i) est plus grande que la valeur de seuil Ct. La frontière entre deux segments est établie de la manière suivante:

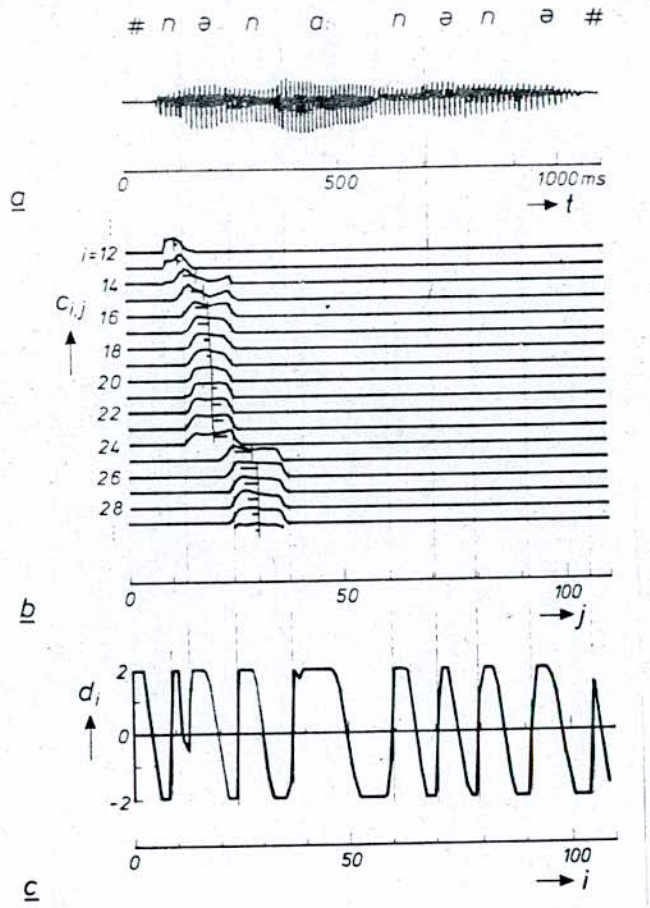


Figure III-4 Résultats de la première méthode
a- Le signal
b- Les fonctions de corrélation
c- La fonction distance

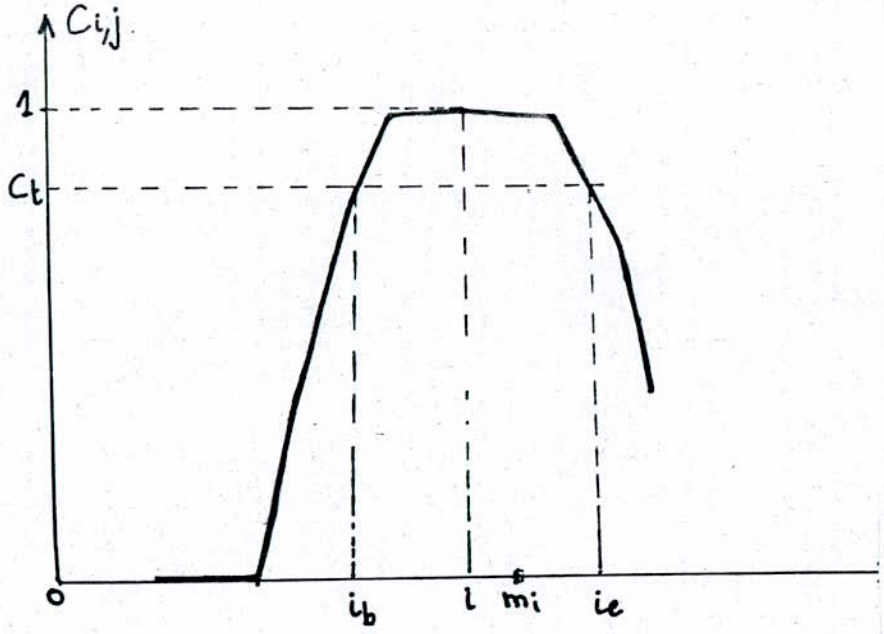


Figure III-3 La corrélation pour le i(ème) spectre

Pour chaque spectre on détermine le centre de masse "mi" qui est défini par:

$$mi = \frac{\sum_{j=ib}^{ie} j\,(Ci,j - Ct)}{\sum_{j=ib}^{ie} (Ci,j - Ct)} \qquad (4)$$

La distance di du spectre au centre de masse mi d'un segment est utilisée pour établir les frontières.

$$di = mi - i \qquad (5)$$

La façon dont ceci est fait exactement peut-être vu en se référant à l'exemple donné (fig III-4).

Pour notre mot "nenaanene" nous avons une représentation du signal de parole, de la corrélation Ci,j des spectres, et la fonction de distance di. Pour chaque i la courbe Ci,j donne la corrélation entre le spectre (i) et les autres spectres (j) dans cette occurence de la parole. De i=14 à i=24 la courbe de corrélation change légèrement. Prés de i=25 on trouve une transition, et ensuite le maximum de Ci,j se déplace vers les valeurs les plus grandes de (j). Le centre de masse

'mi' de la courbe est indiqué par une courte ligne verticale. Les courtes lignes horizontales indiquent la distance (di) du spectre (i) au centre de masse de la courbe de corrélation. Cette distance est aussi donnée à la fin de la figure. Les intersections de l'axe des zéros avec la fonction de distance de plus (+) vers le moins (-) sont toutes proches d'un segment d'état stable. Les intersections de l'axe des zéros avec la courbe de distance de moins (-) vers (+) sont toutes proches d'une zone de transition, et sont considérées comme des frontières entre deux segments d'état stable. La n(éme) frontière déterminée par cette méthode est désignée par g (1,n) (nous reviendrons à celà plus tard).

On peut déterminer les frontiéres des phonèmes pour un certain nombre de mots par cette méthode. Une comparaison de ces résultats avec les résultats obtenus non automatiquement montre que les frontières des phonèmes ont été assez précisement déterminées. Cependant, il peut arriver que certaines frontières ne soient pas trouvées, ou que certains phonèmes soient divisés incorrectement en plus d'un segment. Le nombre de frontières déterminées incorrectement est cependant si petit que cette méthode peut-être utilisée dans un programme de segmentation automatique.

Ces parties du signal ne sont pas réellement des états stables, la forme des spectres change mais pas autant que le changement relatif entre différents segments.

III-3-2 Segmentation avec l'aide d'un modèle de référence

Dans la segmentation en segments d'état stable, tel que discuté dans le sous chapitre précédant, nous n'avons pas établi l'identification phonétique des différents segments. Nous savons cependant combien de phonèmes il y'a dans le mot qu'on doit segmenter, et quels sont ces phonèmes. Nous utilisons cette information pour identifier les segments d'état stable. Ceci peut-être fait par exemple, en connectant les segments d'état stable successifs avec les phonèmes dans la séquence qui intervient dans le mot à segmenter. Cependant ceci ne marche pas toujours, il y'a certains phonèmes (comme les diphtongues /ei/; /ou/; /ai/, et les plosives /b/; /d/; /t/) qui consistent en deux sections d'état plus ou moins stable. De plus des erreurs peuvent-être introduites en délimitant trop de frontières ou pas assez. Il est donc nécessaire à ce stade de s'assurer de la mesure dans laquelle le spectre d'une partie d'état stable correspond réellement au phonème voulu. La procédure que nous suivons est la suivante.

Le mot à segmenter est partagé en états spectraux. La plupart des phonèmes correspondent à un état spectral. Seulement les diphtongues et les plosives correspondent à deux états spectraux. Nous caractérisons chaque état spectral par un spectre qu'on obtient à partir d'une série de spectres correspondant au phonème. Nous appellons ces spectres les

spectres de référence. Ensuite on compare les spectres de l'expression énoncée (les spectres test) avec les spectres de référence de tous les états qui interviennent dans l'expression de l'énoncé. Nous faisons cela de la même façon que nous l'avons décrit dans le sous chapitre précedent, en déterminant la corrélation entre le spectre de référence et le spectre test . Illustrant ceci par un exemple: (fig III-5)

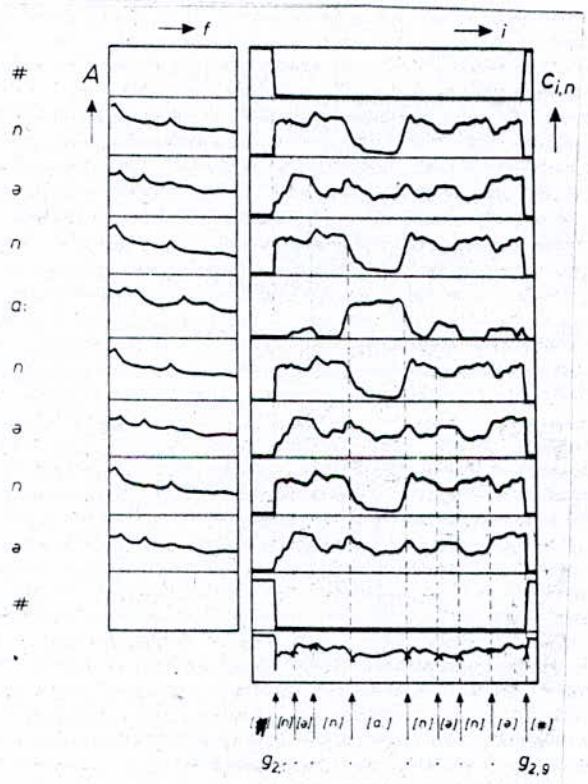


Figure III-5 exemple d'identification phonétique par la seconde méthode

Le mot "nenaanene" mentionné précédemment contient 10 états spectraux: le silence avant le mot /#/, les phonèmes /n/, /e/, /n/, /a/, /n/, /e/, /n/, /e/, et le silence aprés le mot /#/. La figure montre les spectres de références pour chacun de ces états spectraux, et la corrélation Ci,n de tout les spectres test (i) avec les différents spectres de référence (n). La corrélation est élevée si le spectre test ressemble au spectre de référence. La corrélation de tous les spectres test (i) avec les spectres de référence (n) montre, selon le nombre de fois que l'état spectral apparaît, un nombre de régions où la corrélation est élevée. La courbe de corrélation du spectre de référence de /n/ par exemple a quatre régions de ce type; une pour chaque apparition du /n/ dans le mot.

En principe, cette information spectrale peut-être utilisée comme la base pour segmenter un énoncé. La variabilité spectrale de la parole est tellement grande que la corrélation entre le spectre test et le spectre de référence peut toujours être plus faible que la corrélation avec d'autres spectres de référence, qui peuvent donner un état spectral correspondant à une durée nulle. Pour éviter ceci, des restrictions pourraient s'imposer sur la durée de chaque état spectral. La durée du minimum et la durée du maximum de chaque état spectral sont estimées par la mesure de la durée d'un certain nombre de réalisations de cet état spectral.

Nous pouvons maintenant formuler une règle pour la segmentation. Le mot à segmenter contient N états spectraux et il est disponible sous forme de I états spectraux. Le dernier spectre test associé avec le n(éme) état spectral est appelé g(n). Les frontières g(2,n) pour n=1 à N-1 sont les paramètres libres qui doivent-être déterminés. La frontière g(2,0) est au début de l'énoncé et la frontière g(2,n) est positionnée à la fin (g(2,0)=0 et g(2,n)=I). Le paramètre libre doit-être determiné d'une telle manière que la ressemblance entre le spectre de référence et le spectre test soit la plus proche possible, ceci à condition que la durée de chaque état spectral doit-être comprise entre le maximum et le minimum. Cette condition est formulée par:

$$\max(n) = \text{Max}_{g(2,n)} \sum_{i=g(2,n-1)+1}^{g(n)} Ci,n \qquad (6)$$

$$\min(n) = \text{Min}_{g(2,n)} \sum_{i=g(2,n-1)+1}^{g(n)} Ci,n$$

avec la condition $\max(n) \geq g(2,n)-g(2,n-1) \geq \min(n)$

Le résultat de ces calculs peut-être trouvé dans le dernier graphe de la figure III-5. Il montre une partie des courbes de corrélation entre les frontières calculées comme décrit ci-dessus.

En général, la précision pour la détermination des frontières par cette méthode est trop grande pour notre objectif.

III-3-3 Combinaison des deux méthodes de ségmentation

	Avantages	Inconvénients
Première méthode	Les frontières sont déterminées d'une manière précise	Le nombre de frontières est incorrect Pas d'identification phonétique
Deuxième méthode	Le nombre de frontières est correct L'identification phonétique incluse	Les frontières sont localisées incorrectement

Figure III-6 Tableau des avantages et des inconvénients des deux méthodes

On peut voir que les deux méthodes sont complémentaires (fig III-6). Il est donc logique de combiner les deux méthodes de telle manière qu'on retient les avantages de chacune, et ignorer leurs inconvénients. Ceci nous permet de déterminer les frontières des segments par la première méthode et l'identification phonétique par la seconde.

Pour ceci, nous avons besoin d'un algorithme pour combiner les deux méthodes de détermination des frontières (qui combine les frontières déterminées par les deux méthodes).Ceci implique que s'il y'avait des frontières en plus, trouvées par la première méthode, seraient exclues, et que les frontières non trouvées seraient inserées.

Les résultats des deux méthodes sont données dans la figure III-7.

La figure III-7-a donne le spectre du mot à segmenter, la fonction distance di est montrée le long de l'axe des temps avec les frontières g(1,n) qui sont déterminées par la première méthode.

Les frontières déterminées par la seconde méthode g(2,n) sont données par la figure III-7-b, cette figure donne seulement le spectre de référence utilisé pour la détermination

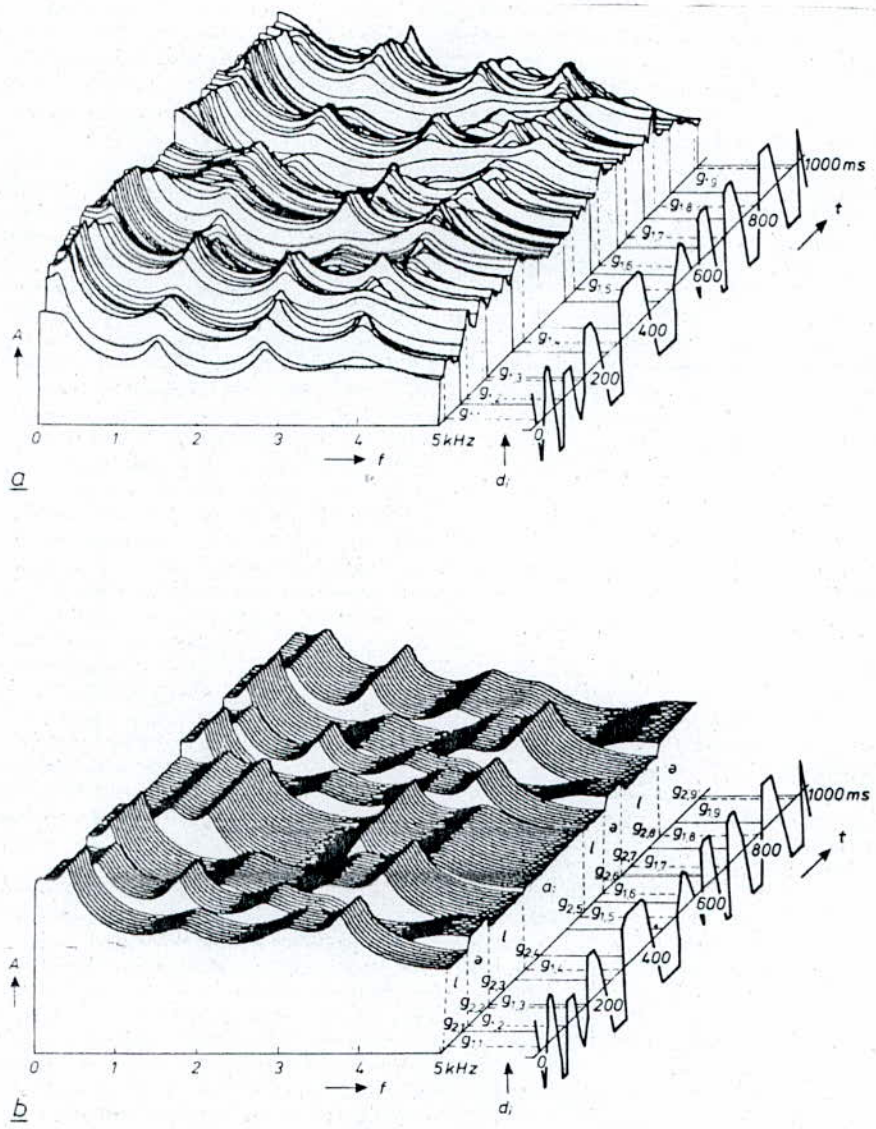
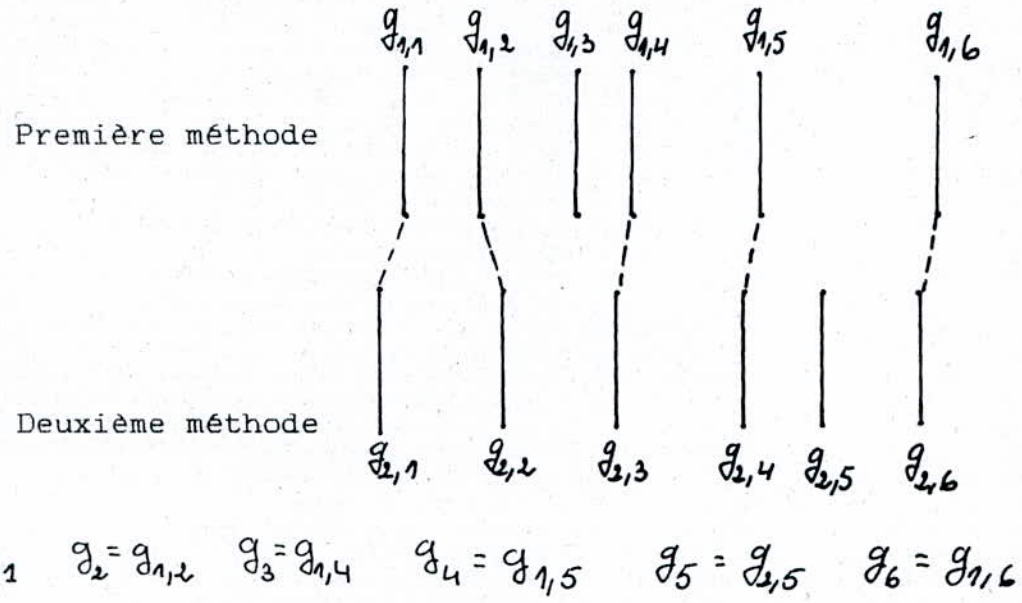


Figure III-7 Résultats des deux méthodes pour le mot "lelaalele"
a/ la première méthode
b/ la seconde méthode

des différentes frontières. Il faut noter qu'il y'a une différence considérable entre le spectre réel (fig III-7-a) et l'état spectral utilisé par la seconde méthode (fig III-7-b). La fonction distance di et les frontières g(2,n) qui en découlent sont données le long de l'axe des temps.

A certains endroits les frontières g(1,n) et g(2,n) sont différentes; les frontières les plus proches l'une de l'autre sont reliées. La localisation finale des frontières découle de la première méthode et l'identification phonétique de la seconde. Le lien entre deux frontières est représenté par une ligne en pointier (fig III-8). Cette procédure de liaison des frontières peut-être continuée jusqu'à un certain nombre de frontières, reste celles qui ne peuvent pas être reliées sans intersection avec les lignes en pointier, les lignes en pointier ne doivent pas être coupées parce que la séquence dans laquelle les segments sont trouvés est invariante. Les frontières g(1,n) qui ne peuvent pas être reliées sont exclues, par contre les frontières g(2,n) qui ne peuvent pas être reliées sont retenues. La localisation finale de ces frontières découle de la seconde méthode.



$g_1 = g_{1,1}$ $g_2 = g_{1,2}$ $g_3 = g_{1,4}$ $g_4 = g_{1,5}$ $g_5 = g_{2,5}$ $g_6 = g_{1,6}$

Figure III-8 Exemple de combinaison des frontières des deux méthodes

III-4 REGLES DE SEGMENTATION DE LA PAROLE EN DIPHONES

La méthode de segmentation en diphones que nous avons choisi, consiste à localiser les frontières ou les limites des diphones à partir de la localisation des frontières des phonèmes. Les règles pour positionner les limites des diphones sont différentes pour chaque phonème.

Pour les voyelles, la limite des diphones est localisée à un temps prédéterminé aprés chaque début de voyelle, la raison pour cela, est qu'une voyelle dans le langage naturel n'est pas affectée ou très peu par la consonne la précedant immédiatemment, mais elle est affectée par la consonne la suivant immédiatement.

Ce temps prédéterminé, varie suivant que l'articulation est rapide ou lente, et selon le locuteur. Les voyelles de la langue française peuvent-être classées en 4 groupes:

-les voyelles fermées : /i/ , /y/ , /u/

-les voyelles mi-fermées : /e/ , /œ/ , /ø̃/ , /ə/ , /o/

-les voyelles mi-ouvertes : /ɔ/ , /ɔ̃/ , /ɑ/ , /ɑ̃/ , /ɛ/ , /ɛ̃/

-les voyelles ouvertes : /a/ , /ɸ/ .

Pour chaque groupe de voyelles le temps est fixe. Par exemple pour le groupe des fermées le temps est égal à 20 ms.

Pour les plosives, /b/, /d/, /t/, /p/, /k/, /g/, la limite du diphone est localisée immédiatemment avant l'explosion.

Pour toutes les autres consonnes la limite est positionnée au milieu du phonème.

III-5 CONCLUSION

Les méthodes de segmentation discutées peuvent être programmées sur ordinateur. L'avantage essentiel de ces méthodes est qu'elles sont indépendentes de la langue et du locuteur, c-à-d qu'elles peuvent être utilisées pour la construction des bibliothèques de diphones pour plusieurs langues et plusieurs locuteurs.

C H A P I T R E IV

P R E D I C T I O N D E S D U R E E S

S E G M E N T A L E S

IV-1 INTRODUCTION

L'objectif de ce chapitre est la définition d'un ensemble de règles pour prévoir les durées des phonèmes.

Un tel modèle est nécessaire pour les systèmes de synthèse par règles, il peut également être utilisé dans les systèmes de synthèse par diphones.

IV-2 MESURE DES DUREES SEGMENTALES

Ces mesures ont été faites dans les laboratoires du CNET. Des enregitrements d'un simple texte lu normallement par 20 locuteurs (12 hommes et 8 femmes), sont analysés en terme du taux d'articulation, le nombre de syllabes prononcées par unité de temps (les pauses exclues). Sur la base de ces mesures, 4 locuteurs sont sélectionnés pour la couverture optimale du taux d'articulation. Leurs taux d'articulation sont respectivement 4.2 syl/s , 4.8 syl/s 5.6 syl/s et 6.1 syl/s, le locuteur dont le taux est le plus élevé (6.1 syl/s) est désigné comme locuteur n° 1 et le locuteur avec le plus bas taux comme locuteur n° 2 .

Les mesures sont effectuées sur des mots monosyllabiques sans aucun sens. L'utilisation de ces mots est imposée par le besoin d'avoir tous les phonèmes dans tous les contextes. Le premier corpus

est constitué de mots CVC et VCV ou V est une voyelle et C une consonne. Le corpus a été lu par les quatre locuteurs déjà choisis.

Pour chaque locuteur, la durée "inhérente" de chaque phonème est définie comme la moyenne des durées du segment mesurées dans tous les contextes. Le plus large écart est de 35 ms pour les voyelles, 18 ms pour les consonnes et 25 ms pour les semi voyelles. L'écart relativement petit pour les consonnes montre une faible influence des voyelles sur les consonnes, mais l'écart relativement large pour les voyelles comme une fonction du contexte des consonnes montre une grande influence des consonnes suivantes sur les voyelles précedentes.

En comparant les durées inhérentes des 4 locuteurs, on a observé un écart atteignant un maximum de 23 ms pour les voyelles et 21 ms pour les consonnes. Les écarts standards relativement petits ont permis le calcul des durées intrinsèques, comme la moyenne des durées inhérentes des 4 locuteurs, pour chaque phonème (fig IV-1).

Ces durées intrinsèques constituent la base des règles de durées du locuteur indépendent. Les coefficients "cointrinsèques" m_c sont calculés pour chaque consonne, afin que le produit de la durée intrinsèque de la voyelle par le coefficient co-intrinsèque spécifique donne la durée éxacte de la voyelle. Il y a six groupes de consonnes à effet égal sur la durée des voyelles (fig IV-2).

a	ɛ	e	i	y	u	o	ɔ	œ	ø	ə	ɛ̃	ɔ̃
177	175	180	170	167	170	186	170	185	186	130	200	200

ɑ̃	ɥ	j	w	f	s	ʃ	v	z	ʒ	p	t	k
200	150	150	144	239	254	250	150	163	173	193	210	210

b	d	g	m	n	l	R
174	167	163	167	157	134	132

Figure VI-1 Les durées intrinsèques en ms mesurées en logatomes

consonnes	p,t,k	b,d,g	v,ʒ,z,R	f,s,ʃ	w,j,l	m,n
locuteur 1	0.85	0.93	1.25	0.88	0.82	0.83
locuteur 2	0.95	1.02	1.56	1.08	1.05	0.95

Figure IV-2 Les coefficients co-intrinsèques pour les deux locuteurs extrêmes

Le deuxième corpus consistait en deux paragraphes avec une durée approximative de 1 mn. Ces deux textes sont lus par les deux locuteurs extrêmes. Les mesures des durées segmentales ont montré de grandes variations en fonction de la position du segment dans un

mot et la position du mot dans la phrase. Pour décrire l'environnement d'un segment d'une manière précise, des informations syntaxico-prosodiques sont utilisées. Ces informations sont décrites à travers un ensemble d'indicateurs qui sont définis comme suit:

-La position d'un mot dans une phrase,

-La position d'un segment dans un mot,

-Présence d'une pause longue (> 100 ms),

-Présence d'une pause finale,

-Présence d'une pause non finale,

-Présence d'une pause courte (< 100 ms)

Le troisième corpus consiste en la mesure des durées segmentales dans les clusters. Un cluster est défini comme deux (ou plus) consonnes ou semi-voyelle adjacentes. Globalement, les consonnes sont plus courtes dans les clusters. Ceci à l'exception des "liquides" qui sont plus longues dans les clusters que dans les milieux des mots plurisyllabiques. La durée des consonnes dans les clusters varie suivant la position du cluster dans le mot et la position du mot dans la phrase. En plus le degré de raccoursissement (d'allongement) dépend de la nature de la consonne.

IV-3 PREDICTION DES DUREES SEGMENTALES

Sur la base de nos mesures, on suggère que la durée segmentale peut-être prédite par:

La durée de la voyelle = DI . Vi

La durée de la consonne = DI . Cij

où DI est la durée intrinsèque d'un segment et Vi et Cij sont des coefficients de raccourcissement ou d'allongement dépendant de la position d'un segment dans un mot ou une phrase "i" . Dans le cas des consonnes, ces coefficients sont également dépendants de l'appartenance à la classe phonétique "j" de la consonne. La prédiction de la durée des voyelles et des consonnes est donnée respectivement par les figures IV-3 et IV-4.

IV-3-1 Paramètres de prédiction

Les valeurs des coefficients Vi, Cij sont déterminées sur la base des mesures acoustiques. Pour chaque apparition d'un phonème dans une position donnée, la valeur du coefficient correspondant est calculée pour obtenir le meilleur "marqueur" entre les durées observées et les durées prédites. Chaque coefficient Vi, Cij peut prendre deux valeurs numériques pour prédire le système de durée pour les locuteurs 1 et 2 (figure IV-5 et IV-6).

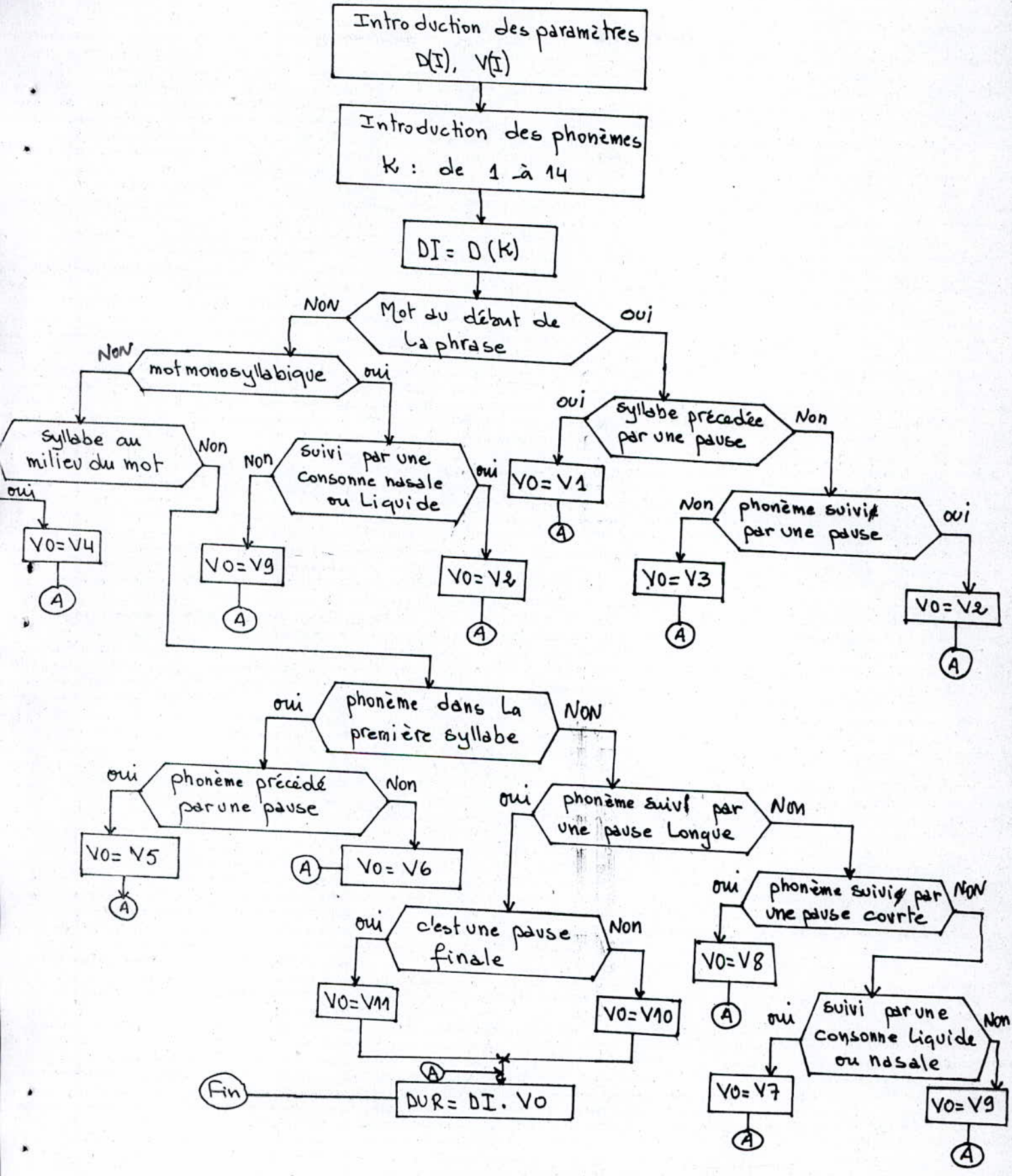
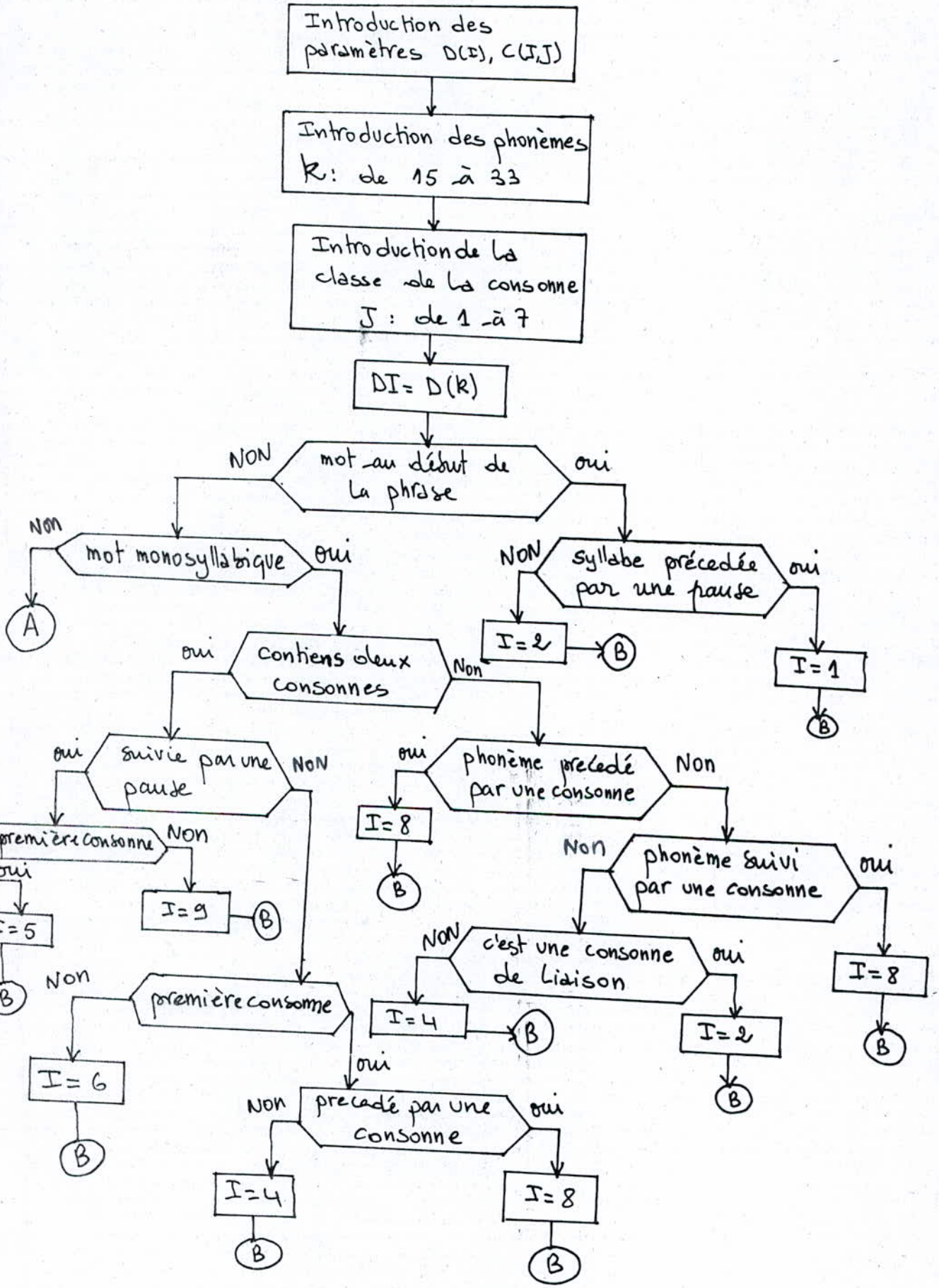


Figure IV-3 Organigramme de prédiction des durées pour les voyelles


Introduction des paramètres D(I), C(I,J)
Introduction des phonèmes
K: de 15 à 33
Introduction de la classe de la consonne
J : de 1 à 7
DI = D(K)
mot au début de la phrase
NON
oui
mot monosyllabique
Non
oui
A
syllabe précedée par une pause
NON
oui
I=2
B
I=1
B
contiens deux consonnes
oui
Non
suivie par une pause
oui
NON
première consonne
oui
Non
I=5
B
I=9
B
phonème precedé par une consonne
oui
Non
I=8
B
phonème suivi par une consonne
Non
oui
c'est une consonne de liaison
NON
oui
I=4
B
I=2
B
I=8
B
première consomne
Non
oui
I=6
B
precadé par une consonne
Non
oui
I=4
B
I=8
B

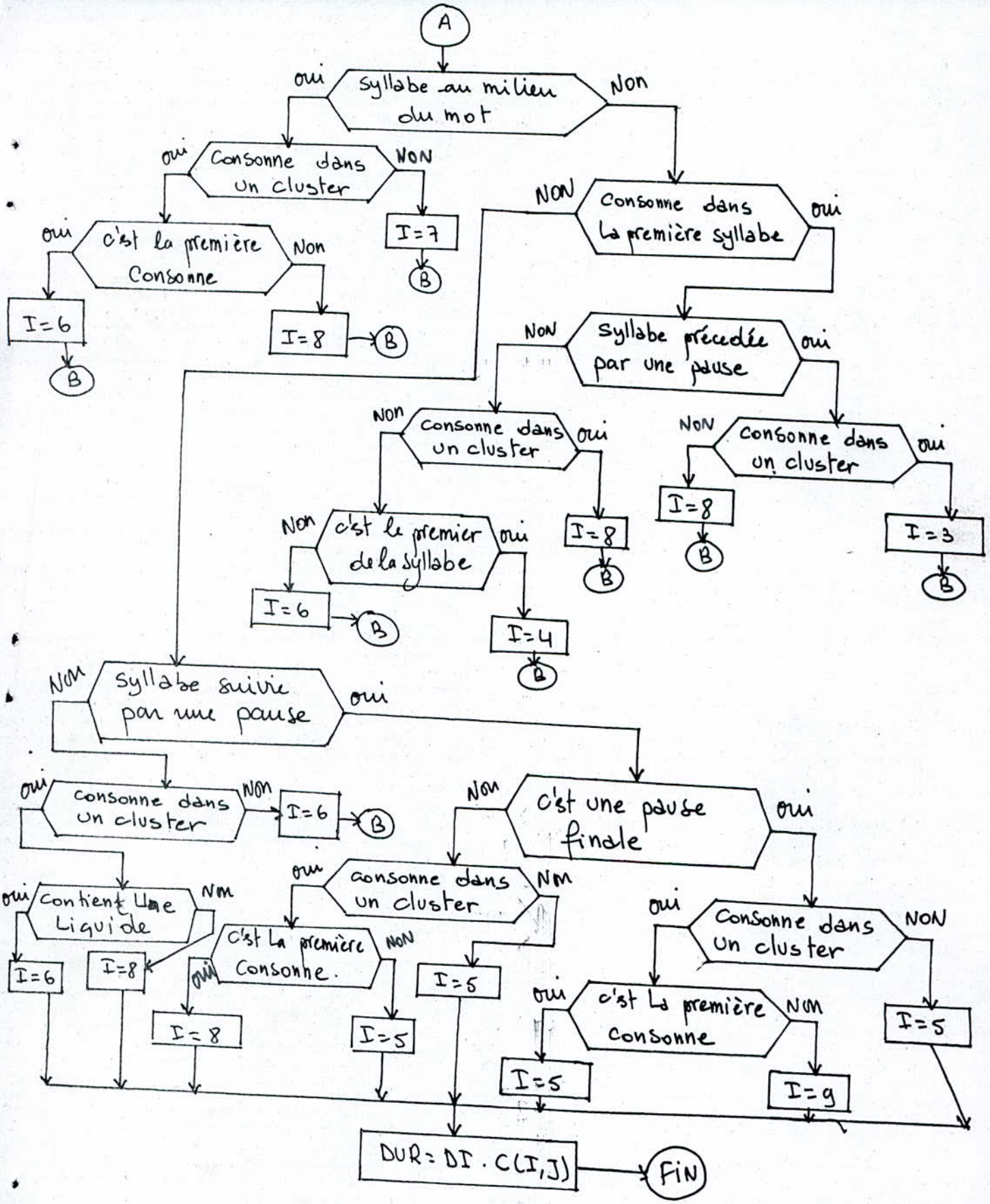


Figure IV-4 Organigramme de prédiction des durées pour les consonnes

coefficients	V1	V2	V3	V4	V5	V6
locuteur 1	0.36	0.36	0.86	0.40	0.45	0.38
locuteur 2	0.56	0.38	1.10	0.40	0.59	0.43

coefficients	V7	V8	V9	V10	V11
locuteur 1	0.43	0.66	0.52	0.86	0.66
locuteur 2	0.45	0.80	0.60	1.10	0.80

Figure IV-5 Valeurs des coefficients Vi

i	j	C1j	C2j	C3j	C4j	C5j	C6j	C7j	C8j	C9j
b,d,g	1	0.52	0.32	0.56	0.34	0.60	0.40	0.38	0.30	0.59
p,t,k	2	0.57	0.36	0.57	0.47	0.62	0.50 0.57	0.49	0.33	0.80 0.95
v,z,ʒ	3	0.50	0.42	0.52	0.42 0.48	0.55 0.60	0.49	0.46	0.33 0.38	0.80 0.90
f,s,ʃ	4	0.48	0.40	0.52	0.48	0.60	0.53	0.44 0.51	0.24 0.32	0.90
m,n	5	0.40	0.30	0.45	0.40	0.50	0.42	0.40	0.34	0.50
R,l	6	0.37 0.57	0.30	0.45 0.57	0.32 0.38	0.46 0.58	0.33	0.31	0.45	0.75
w,j,y	7	0.30	0.25	0.33 0.40	0.26	0.60	0.33	0.40	0.25	0.80 0.86

Figure IV-6 Les coefficients Cij (i:classe du coefficient et j la classe de la consonne),pour les cases à deux valeurs la 1(ére) correspond au locuteur 1 la seconde au locuteur 2.

IV-3-2 Validation du modèle

Le modèle fournit un ensemble de règles pour prévoir les durées segmentales des différents phonèmes dans un mot ou une phrase. Ce modèle a été testé sur une phrase lue par les deux locuteurs extrêmes (fig IV-7). Afin de valider le modèle d'autres mesures ont été faites.

Premièrement, les deux locuteurs ont fait chacun un nouveau enregistrement des textes qui ont servi à l'élaboration du modèle. Les durées segmentales mesurées étaient différentes des durées prédites, un écart standard a été observé (fig IV-8).

Deuxièment, chacun des deux locuteurs a fait deux autres enregistrements d'un nouveau texte. Une comparaison entre les durées mesurées et les durées prédites a fourni de nouvelles valeurs d'écart standard (fig IV-9)

Le même ordre de grandeur des écarts standards des figures IV-8 et IV-9 , a montré que le modèle est acceptable.

	Consonnes	voyelles
Locuteur 1	13 ms	12 ms
Locuteur 2	11 ms	10 ms

Figure IV-8 Ecart standard du premier essai

Phonèmes	l	a	v	i	s	e	k	y	l	o	ʁ
Coefficients	C1	V1	C4	V9	C2	V6	C6	V7	C6	V2	C4
Dur mes n°1	50	45	80	70	90	50	120	70	50	60	40
Dur pré n°1	49	63	63	88	101	68	105	73	44	67	42
Dur mes n°2	60	100	90	100	110	80	120	76	50	90	40
Dur pré n°2	76	99	72	110	101	77	119	85	44	71	50

Phonèmes	a	l	ɑ̃	t	i	a	l	ɛ̃	t	ɛ	ʁ
Coefficients	V6	C7	V4	C6	V9	V2	C2	V6	C7	V4	C6
Dur mes n°1	70	40	90	140	70	50	40	80	90	50	40
Dur pré n°1	67	41	80	105	88	63	40	76	102	69	43
Dur mes n°2	60	40	80	100	120	80	40	70	90	50	40
Dur pré n°2	76	41	100	119	102	68	40	86	102	69	43

Phonèmes	ʒ	œ	ʁ	d	e	m	ɛ	z	ɔ̃
Coefficients	C8	V7	C8	C8	V2	C4	V6	C5	V11
Dur mes n°1	50	55	40	50	50	60	80	70	140
Dur pre n°1	37	79	59	50	65	53	66	89	132
Dur mes n°2	40	70	40	50	60	70	70	100	150
Dur pré n°2	37	83	59	50	69	66	75	97	160

Figure IV-7 Prédiction des durées segmentales pour la phrase "La vie s'écoule au ralenti à l'intérieur des maisons"

Dur mes: Durées mesurées (en ms)
Dur pré: Durées prédites (en ms)

	Consonnes	Voyelles
Locuteur 1	17 ms	15 ms
	17 ms	15.5 ms
Locuteur 2	15 ms	15 ms
	15 ms	14 ms

Figure IV-9 Ecarts standards du deuxième essai

IV-4 COMPARAISON AVEC D'AUTRES MODELES

Le modèle de règles décrit ici ne devrait pas être considéré comme inclusif. Il omet beaucoup de petites variations telles que celles affectant les consonnes dans les clusters, et les durées de fermetures et d'éclatements pour les plosives.

En général, la structure analytique de ce modèle de règles est grandement influencée par celle de KLATT. La différence entre le modèle de KLATT et celui-ci est que les règles de KLATT limitent le "raccoursissement" des durées segmentales à une durée minimale fixée à priori. Une telle limite, qui est soigneusement mesurée par KLATT pour les segments accentués en anglais, n'est pas utile en français, où la "position accent" n'est pas si définie.

Les différences sont plus grandes avec le modèle de O'SHAUGHNESSY (1984). Les règles de O'SHAUGHNESSY utilisent toujours les durées segmentales de base qui sont augmentées ou réduites afin de prendre le contexte en considération. Mais les valeurs numériques utilisées dans ce modèle sont derivées de la moyenne des durées segmentales prise sur 29 locuteurs mesurées sur un seul corpus.

Un tel corpus est trop petit pour permettre le calcul des durées de base, et la mesure des modifications dûes au contextes. En géneral la prédiction des durées sont plus ou moins similaire avec notre modèle, mais les règles de O'SHAUGHNESSY sont beaucoup plus compliquées.

IV-5 CONCLUSION

L'étude d'un tel modèle de prédiction des durées segmentales, nous permet de calculer les variations des durées des phonèmes dans différents contextes. Cette étude rentre dans le niveau prosodique du traitement automatique de la parole.

Ce modèle a plusieurs limites, surtout au niveau des consonnes et des voyelles nasales. Les problèmes sont plus compliqués dans la situation où on a une consonne nasale suivie ou précedée par une voyelles nasale. On peut voir que les règles qui régissent ces situations n'existent pas dans notre programme.

C O N C L U S I O N G E N E R A L E

Dans ce projet, nous avons étudié quelques notions sur la production naturelle de la parole et le schéma général de la communication parlée homme-machine, en mettant l'accent sur les différentes possibilités de division du signal de parole. Cette étude nous a permis de comprendre l'interêt de la segmentation qui est une phase importante dans le traitement automatique de la parole.

Pour faire la segmentation, une opération qui s'effectue au niveau acoustique, une analyse du signal de parole est nécessaire:

-Un prétraitement, qui consiste en la transformation des variations de pressions de l'air en signal électrique, un filtrage, un fenêtrage, un échantillonage,etc...

-Une analyse spectrale du signal, différentes méthodes sont utilisées: méthodes de prédiction linéaire, analyse de FOURIER, etc...

Nous avons étudié en détail trois méthodes de segmentation:

La méthode de segmentation en segments d'état stable, la segmentation avec l'aide d'un modèle de référence, et la segmentation en diphones. Une comparaison des avantages et incovénients d'application propre à chacune des deux premières

méthodes, nous a imposé la combinaison de ces deux dernières de telle manière à retenir les avantages de chacune d'elles, et ignorer leurs inconvénients, ce qui nous a permis d'avoir une segmentation avec le minimum d'erreurs. Mais même ainsi, des erreurs peuvent apparaitre, surtout sur la position des frontières lorsque la combinaison n'est pas possible. Des erreurs peuvent aussi être introduites par le module d'analyse, les conditions d'enregistrements,etc...

En ce qui concerne la prédiction des durées segmentales, nous avons étudié un modèle qui est utilisé dans les modules prosodiques des systèmes de synthèse. Ce modèle qui développé sur la base des mesures des durées observées des différents segments dans différents contextes. Les erreurs qu'on peut rencontrer dans ce modèle sont dues aux erreurs de segmentation. L'erreur relative observée montre que le modèle est acceptable, et une comparaison de ce modèle avec d'autres montre que celui-ci n'est pas aussi mauvais que les autres.

Néanmoins, notre étude pourrait permettre d'éffectuer des mesures sur les valeurs numériques et la constitution d'un modèle de référence pour la langue arabe, car les méthodes de segmentation discutées dans ce projet sont indépendantes de la langue. Cela peut-être l'objet d'une recherche ultérieure bien approfondie.

A N N E X E

Nous présentons en annexe le programme de prédiction des durées segmentales. L'utilisation de ce programme passe par plusieurs étapes:

-La phrase à traiter doit être traduite en écriture phonétique, en respectant les règles phonologiques de la langue.

-La phrase doit être découpée en mots,

-Chaque mot doit être découpé en syllabes,

-Chaque syllabe doit être découpée en phonèmes,

-On doit déterminer toute les pauses (courtes et longues) .

Le programme utilisé nous pose des questions auxquelles nous devons répondre par "oui" ou par "non". Lorsque la réponse à la question est "oui" nous tapons "1", par contre lorsque la la réponse est "non" nous tapons "0".

Pour l'introduction des différents phonèmes on utilise les indices de 1 à 33 (tableau A-1), pour la classe des consonnes nous utilisons les indices de 1 à 7 (tableau A-2).

Les consonnes	b,d,g	p,t,k	v,ʒ,z	f,ʃ,s	m,n	R,l	w,j,ɥ
La classe	1	2	3	4	5	6	7

Tableau A-2 Classes des consonnes

Les phonèmes	a	ɛ	e	i	y	u	o	ɔ	oe	ø	ə	ɛ̃
La classe	1	2	3	4	5	6	7	8	9	10	11	12

Les phonèmes	ɔ̃	ɑ̃	ɥ	j	w	f	s	ʃ	v	z
La classe	13	14	15	16	17	18	19	20	21	22

Les phonèmes	ʒ	p	t	k	b	d	g	m	n	l	R
La classe	23	24	25	26	27	28	29	30	31	32	33

Tableau A-1 Les phonèmes

Le tableau A-1 comporte 33 indices:

-Les indices de 1 à 14 corresponds aux voyelles

-Les indices de 18 à 33 corresponds aux consonnes, les semi-voyelles ont leurs indices de 15 à 17.

```
0 '*****************************************************************
0 '        PREDICTION  DES  DUREES  SEGMENTALES
0 '*****************************************************************
0 '
0 '     INTRODUCTION DES PARAMETRES DEPENDANTS DU LOCUTEUR
0 '
5 DEFINT D
0 DIM V(111),C(90,70),D(133),DUR(100)
0 CLS
2 OPEN "I",#1,"VI"
4 OPEN "I",#2,"CIJ"
0 '
00 '              INTRODUCTION   DES  PARAMETRES   V(I)
10 '
20 FOR I=1 TO 11
30 INPUT #1,V(I)
40 NEXT I
50 '
60 CLS
70 '
80 '              INTRODUCTION   DES   PARAMETRES   C(I,J)
90 '
00 FOR I=1 TO 9
10 FOR J=1 TO 7
20 INPUT #2, C(I,J)
30 NEXT J
40 NEXT I
50 '
52 '
54 '          INTRODUCTION DES PARAMETRES INDEPENDANTS DU LOCUTEUR
56 '
60 '
40 '            INTRODUCTION   DES   DUREES  INTRINSEQUES
50 '
60 D(1)=177 : D(2)=175 : D(3)=180 : D(4)=170 : D(5)=167 : D(6)=170 :
   D(8)=170 : D(9)=185 : D(10)=186: D(11)=130: D(12)=200: D(13)=200:
   D(15)=144: D(16)=150: D(17)=150: D(18)=239: D(19)=254: D(20)=250:
   D(22)=163
65 D(23)=173 : D(24)=193 : D(25)=210 : D(26)=210 : D(27)= 174 :
   D(29)=163 : D(30)=167 : D(31)=157 : D(32)=134 : D(33)= 132
68 D(7)=186:  D(14)=200:  D(21)=150: D(28)=167
70 '
80 CLS
90 '
00 '          INTRODUCTION  DU  NOMBRE  DE  PHONEMES
10 '
20 INPUT"COMBIEN  DE  PHONEMES VOUS  AVEZ  DANS  LA PHRASE : A";A
30 FOR B=1 TO A
40 CLS
50 IF B>1 THEN 510
60 '
70 '          INTRODUCTION  DES   PHONEMES
80 '
90 INPUT"QUEL EST VOTRE PREMIER PHONEME DE 1 A 33 :K";K
00 GOTO 520
```

```
510 INPUT"QUEL EST LE PHONEME SUIVANT DE 1 A 33 :K";K
520 DI(B)=D(K)
530 IF K>14  THEN 870
532 '
534 '  CALCUL DES DUREES SEGMENTALES POUR LES VOYELLES
536 '
540 INPUT"LE MOT EST AU DEBUT DE LA PHRASE : H1";H1
550 IF H1=0 THEN 621
560 INPUT"LA SYLLABE EST PRECEDE PAR UNE PAUSE : H2";H2
570 IF H2=0  THEN 590
580 V0=V(1) : GOTO 850
590 INPUT"EST-IL SUIVI PAR UNE PAUSE : H3";H3
600 IF H3=0  THEN 620
610 V0=V(3) : GOTO 850
620 V0=V(2) : GOTO 850
621 INPUT"DANS UN MOT MONOSYLLABIQUE  :HA";HA
622 IF HA=0 THEN 630
623 INPUT"SUIVIE PAR UNE CONSONNE NASALE OU LIQUIDE :HB";HB
624 IF HB=0 THEN 626
625 V0=V(2) : GOTO 850
626 V0=V(9) : GOTO 850
630 INPUT"EST-IL DANS UNE SYLLABE AU MILIEU DU MOT:H4";H4
640 IF H4=0 THEN 660
650 V0=V(4) : GOTO 850
660 INPUT"EST-IL DANS LA PREMIERE SYLLABE : H5";H5
670 IF H5=0 THEN 720
680 INPUT"EST-IL PRECEDE PAR UNE PAUSE : H6";H6
690 IF H6=0 THEN 710
700 V0=V(5) : GOTO 850
710 V0=V(6) : GOTO 850
720 INPUT"EST-IL SUIVI PAR UNE PAUSE LONGUE :H7";H7
730 IF H7=0 THEN 760
733 INPUT"C'EST UNE PAUSE FINALE :H8";H8
736 IF H8=0 THEN 750
740 V0=V(11) : GOTO 850
750 V0=V(10) : GOTO 850
760 INPUT"EST-IL SUIVI PAR UNE PAUSE COURTE :H9";H9
790 IF H9=0 THEN 810
800 V0=V(8) : GOTO 850
810 INPUT"EST-IL SUIVI PAR UNE CONSONNE LIQUIDE:H10";H10
820 IF H10=0 THEN 840
830 V0=V(7) : GOTO 850
840 V0=V(9) : GOTO 850
850 DUR(B)=DI(B)*V0
860 GOTO 1710
862 '
864 ' CALCUL DES DUREES SEGMENTALES POUR LES CONSONNES
866 '
870 INPUT"QUELLE EST LA CLASSE DE VOTRE CONSONNE DE 1 A 7:J";J
880 INPUT"LE MOT EST-IL AU DEBUT DE LA PHRASE:P1";P1
890 IF P1=0 THEN 940
900 INPUT"LA SYLLABE EST PRECEDE PAR UNE PAUSE:P2";P2
910 IF P2=1 THEN 930
920 I=2 : GOTO 1700
930 I=1 : GOTO 1700
```

```
940 INPUT"LE MOT EST MONOSYLLABIQUE:P3";P3
950 IF P3=0 THEN 1210
960 INPUT"MOT CONTENANT DEUX CONSONNES: P4";P4
970 IF P4=0 THEN 1110
980 INPUT"MOT SUIVI PAR UNE PAUSE:P5";P5
990 IF P5=0 THEN 1040
1000 INPUT"C'EST LA PREMIERE CONSONNE:P6";P6
1010 IF P6=0 THEN 1030
1020 I=5 :GOTO 1700
1030 I=9 :GOTO 1700
1040 INPUT"C'EST LA PREMIERE CONSONNE:P7";P7
1050 IF P7=0 THEN 1100
1060 INPUT"SYLLABE PRECEDEE PAR UNE CONSONNE:P8";P8
1070 IF P8=0 THEN 1090
1080 I=8 : GOTO 1700
1090 I=4 : GOTO 1700
1100 I=6 : GOTO 1700
1110 INPUT"PHONEME PRECEDE PAR UNE CONSONNE :P9";P9
1120 IF P9=0 THEN 1140
1130 I=8 : GOTO 1700
1140 INPUT"PHONEME SUIVIE PAR UNE CONSONNE:P10";P10
1150 IF P10=0 THEN 1170
1160 I=8 : GOTO 1700
1170 INPUT"C'EST UNE CONSONNE DE LAISON :P11";P11
1180 IF P11=0 THEN 1200
1190 I=2 :GOTO 1700
1200 I=4 : GOTO 1700
1210 INPUT"PHONEME DANS UNE SYLLABE AU MILIEU DU MOT:Q1";Q1
1220 IF Q1=0 THEN  1300
1230 INPUT"EST-IL DANS UN CLUSTER :Q2";Q2
1240 IF Q2=0 THEN 1290
1250 INPUT"C'EST LA PREMIERE CONSONNE :Q3";Q3
1260 IF Q3=0 THEN 1280
1270 I=6 : GOTO 1700
1280 I=8 : GOTO 1700
1290 I=7 : GOTO 1700
1300 INPUT"EST-IL DANS LA PREMIERE SYLLABE  :Q4";Q4
1310 IF Q4=0 THEN 1460
1320 INPUT"EST-IL DANS UNE SYLLABE PRECEDE PAR UNE PAUSE:Q5";Q5
1330 IF Q5=0 THEN 1380
1340 INPUT"EST-IL DANS UN CLUSTER:Q6";Q6
1350 IF Q6=0 THEN 1370
1360 I=3 : GOTO 1700
1370 I=8 : GOTO 1700
1380 INPUT"EST-IL DANS UN CLUSTER:Q7";Q7
1390 IF Q7=0 THEN 1410
1400 I=8 : GOTO 1700
1410 INPUT"EST-IL LE PREMIER DE LA SYLLABE:Q8";Q8
1420 IF Q8=0 THEN 1450
1430 I=4 : GOTO 1700
1450 I=6 : GOTO 1700
1460 INPUT"EST-IL DANS UNE SYLLABE SUIVIE PAR UNE PAUSE:Q9";Q9
1470 IF Q9=0 THEN 1630
1480 INPUT"C'EST UNE PAUSE FINALE :Q10";Q10
1490 IF Q10=0 THEN 1570
1500 INPUT"EST-IL DANS UN CLUSTER :Q11";Q11
```

```
1510 IF Q11=0 THEN 1560
1520 INPUT"C'EST LA PREMIERE CONSONNE :Q12";Q12
1530 IF Q12=0 THEN 1560
1540 I=5 : GOTO 1700
1560 I=9 : GOTO 1700
1570 INPUT"EST-IL DANS UN CLUSTER:Q13";Q13
1580 IF Q13=0 THEN 1620
1590 INPUT"C'EST LA PREMIERE CONSONNE :Q14";Q14
1600 IF Q14=0 THEN 1620
1610 I=8 : GOTO 1700
1620 I=5 : GOTO 1700
1630 INPUT"EST-IL DANS UN CLUSTER :Q15";Q15
1640 IF Q15=0 THEN 1670
1650 INPUT"LE CLUSTER CONTIENT LA LIQUIDE (1):Q16";Q16
1660 IF Q16=0 THEN 1680
1670 I=6 : GOTO 1700
1680 I=8
1700 DUR(B)=DI(B)*C(I,J)
1710 NEXT B
1720 CLS
1730 '
1740 '            AFFICHAGE  DES  RESULTATS
1750 '
1760 LOCATE 1,15 : PRINT"VOICI LES DUREES SEGMENTALES DE VOTRE PHRASE"
1770 FOR B=1 TO A
1780 LOCATE 1+B,15 : PRINT DUR(B)
1790 NEXT B
1800 END
```

B I B L I O G R A P H I E

-BARTKOVA.K, SORIN.C. A model of segmental duration for speech synthesis in French. Speech Communnication n°6 , 1987.

-BARTKOVA.K, SORIN.C. Predictive model of segmental duration. 109 ème meeting, Austin 1984.

-CA TIER.E. La parole: Analyse-Synthèse-Reconnaissance. TLE n° 489, Decembre 1983.

-CRYSTAL.T.H, HOUSE.A.S. Segmental duration in connected speech signal. Jour Acous of Amer n°3, Septembre 1982.

-DELAIRE.F, ROSSI.M. Segmentation et étiquitage pour un système de reconnaissance automatique multilocuteur. 15 ème journées d'étude sur la parole, Aix en Provence, Mai 1986.

-FIMBOT.F, MARCUS.S.M, CHOLET.G. Localisation et représentation temporelle d'évenements phonètiques. 15 éme journées d'étude sur la parole, Aix en Provence, Mai 1986.

-GUIBERT.J. La parole: Compréhension et synthèse par les ordinateurs. Presses Universitaires de France, 1979.

-MICLET.L, VICARD.D. Reconnaissance des parties stebles de parole continue pour le décodage acoustico-phonétique. 15 ème journées d'étude sur la parole, Aix en provence, Mai 1986.

-VAN.HEMERT.J.P. Automatic segmentation of speech into diphone. Philips Technical Review, volume 43 n°9, septembre 1987.