INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
/_ /_ /_ /_ /_ /_ /_ /_ /_ /_ /

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité: « Signal, Image, Parole, Télécoms »

préparée au laboratoire CLIPS-IMAG « Communication Langagière et Interaction Personne-Système »
dans le cadre de **l'Ecole Doctorale *« Electronique, Electrotechnique, Automatique, Télécommunications, Signal »***

présentée et soutenue publiquement

par
NGUYEN Quoc Cuong
le 19 Juin 2002

Titre:

RECONNAISSANCE DE LA PAROLE EN LANGUE VIETNAMIENNE

Directeur de thèse:
Eric CASTELLI

JURY

M.	Pierre-Yves COULON	, Président
Mme.	Régine ANDRE-OBRECHT	, Rapporteur
Mme.	Geneviève CAELEN-HAUMONT	, Rapporteur
Mme.	PHAM THI Ngoc Yen	, Invitée
M.	TRINH Van Loan	, Examinateur
M.	Laurent BESACIER	, Examinateur
M.	Eric CASTELLI	, Directeur de thèse

A mes parents

Remerciements

Ce mémoire de thèse présente les travaux de recherche que j'ai effectués au sein du laboratoire Communication Langagière et Interaction Personne-Système (CLIPS) de l'institut Informatique Mathématique Appliquée de Grenoble (IMAG).

Je voudrai tout d'abord remercier Jean Caelen, directeur de recherches au CNRS, pour m'avoir accueilli au sein du laboratoire CLIPS.

Je tiens à remercier Pierre-Yves Coulon, professeur à l'INPG, qui m'a fait l'honneur d'avoir accepté d'être le président du jury.

Je souhaite remercier Régine André-Obrecht, professeur à l'Université Paul Sabatier, et Geneviève Caelen-Haumont, directeur de recherches au CNRS, pour avoir accepté d'être les rapporteurs de cette thèse. Leurs remarques pertinentes sur le contenu m'ont permis d'améliorer la qualité de ce document.

Je souhaite également remercier Pham Thi Ngoc Yen et Trinh Van Loan, les maîtres de conférences à l'Institut Polytechnique de Hanoi, pour avoir examiné ce travail et aussi pour leurs suggestions sur celui-ci.

Je souhaite exprimer ma gratitude à mon directeur de thèse Eric Castelli, maître de conférences à l'INPG, qui m'a dirigé et m'a guidé tout au long de mes travaux.

Je souhaite également exprimer ma gratitude à Jean-François Serignat, maître de conférences à l'INPG, et surtout, Laurent Besacier, maître de conférences à l'UJF, qui m'ont soutenu pendant l'épreuve difficile de la rédaction finale de ce document, pour leurs précieux conseils et leur aide dévouée.

Je tiens à remercier tous les membres du laboratoire CLIPS, tous mes collègues et amis du laboratoire, Dan Istraté pour les discussions amicales, Dominique Vaufreydaz pour son logiciel EMACOP et tous les autres pour leur soutien.

Je tiens à remercier vivement tous mes amis vietnamiens, Benoît Boutillier et Olivier Hafner, mes amis français, pour leur sympathie et l'aide qu'ils m'ont apportées au cours de ma thèse à Grenoble.

Enfin je remercie mes proches, mes parents, mes sœurs, mes beaux-frères et Tuyet, mon amie, pour leur soutien et leur confiance tout au long de cette épreuve

Table des matières

Table des figures

Liste des tableaux

1

Introduction

La parole constitue le moyen le plus naturel de communiquer entre deux personnes. Néanmoins, avec les progrès de l'électronique et de l'informatique, les machines perfectionnées envahissent le quotidien de l'être humain et celui-ci veut aussi utiliser la communication parlée pour interagir avec ces machines, car la communication est alors plus naturelle.

Cette communication orale permet, en effet, de commander certaines applications complexes plus facilement en libérant les mains et la vue pour d'autres activités. Il est possible, par exemple, de rentrer des données dans un ordinateur par la parole, et de remplacer ainsi la saisie de ces données au clavier ou à la souris, saisie manuelle qui ne peut pas toujours se faire dans des milieux industriels ou dans le cas d'activités nomades. Dans le domaine des télécommunications, il est courant maintenant d'être confronté à des systèmes de répondeurs, sans opérateurs humain et automatisés par l'utilisation de systèmes d'interaction homme-machine fondés sur la communication parlée (répondeurs téléphoniques des opérateurs de téléphones mobiles, interrogation des comptes bancaires, commandes en « direct » dans les sociétés de vente par correspondance etc.). Enfin, la communication parlée semble indispensable dans certaines situations pour suppléer des « canaux de communication » défectueux chez l'être humain en cas de handicap, comme chez les malvoyants [Bellik, 1997].

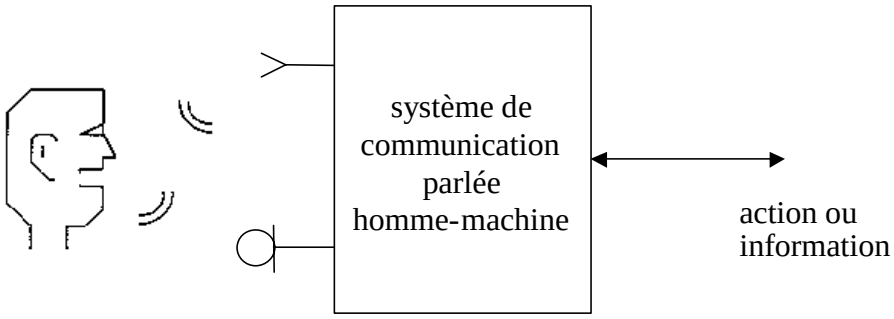


Figure 1.1 Communication parlée homme-machine (d'après Schafer, 1994)

La figure **1.1** présente le diagramme simplifié de la communication parlée homme-machine [Schafer, 1994]. Grâce à un microphone, jouant le rôle du capteur permettant l'acquisition de la voix humaine, et à un haut-parleur (ou des écouteurs) qui restituera la voix synthétique produite par la machine, l'être humain peut communiquer avec le système.

Cependant, pour cela, le système de communication parlée doit numériser puis analyser le canal d'entrée de la voix humaine, puis déterminer alors quelle action il doit réaliser ou quelle information doit être transmise vers d'autres systèmes ou machines. Ce processus complexe demande à la machine des capacités de reconnaissance et de compréhension de la parole. Le système doit également pouvoir produire une voix synthétique en sortie. Celle-ci peut être alors employée pour fournir un "feed-back" pour assurer l'être humain que la machine a correctement compris le discours d'entrée, mais peut également être essentielle pour renvoyer vers l'auditeur des informations, résultats d'une requête ou d'un calcul quelconque.

La reconnaissance de la parole est l'un des modules de l'ensemble complexe implanté dans une machine douée de capacités de communication orale. Dans un sens limité et strict la reconnaissance de la parole peut être considérée comme la conversion d'un signal de parole, formé d'ondes acoustiques, en mots. Elle comprend l'acquisition et la numérisation du signal de parole, puis l'analyse (généralement une analyse spectrale) et la conversion de ce signal en unités élémentaires du discours telles que les phonèmes ou les mots, et, enfin, l'interprétation des séquences converties en ordre afin de permettre la correction des mots reconnus incorrects ou bien leur transmission vers d'autres modules réalisant les autres processus linguistiques comme l'analyse syntaxique et sémantique pour la compréhension de la parole [Juang, 1998].

Pendant les vingt dernières années, les systèmes de reconnaissance de la parole se sont développés rapidement. Ils ont évolué des petits systèmes de reconnaissance de mots isolés

avec des petits dictionnaires à des systèmes beaucoup plus performants capables de reconnaître de la parole continue pour de grands vocabulaires.

Cependant, les systèmes les plus souvent étudiés ont été réalisés essentiellement pour les langues occidentales comme l'anglais, le français, l'allemand, l'italien ou l'espagnol [Barnett et al, 1995]. De nos jours, les activités de recherche sur la reconnaissance de la parole tendent à s'internationaliser, car tous les pays souhaitent posséder la technologie et le savoir faire pour reconnaître automatiquement leur langue maternelle. De plus, avec la mondialisation et les facilités actuelles pour communiquer entre les pays, même si ceux-si sont très éloignés géographiquement, des systèmes multilingues sont de plus en plus développés associant alors des modules de reconnaissance de la parole à des modules de traduction automatique. Au niveau de la recherche des projets internationaux de la traduction multi-langues de la parole sont à l'étude, comme le projet C-STAR [Boitet et al, 1998]

Pour réaliser un système de reconnaissance de la parole, une connaissance des caractéristiques linguistiques et acoustiques de la langue est nécessaire. En effet, chaque langue présente des caractéristiques linguistiques et acoustiques différentes. Cette connaissance permet d'adapter le système aux caractéristiques propres de la langue et ainsi d'améliorer le taux de reconnaissance au final. Elle peut servir également pour l'adaptation de techniques développées, appliquées et validées à une langue pour une autre langue.

Notre travail se concentre sur la reconnaissance automatique de la parole en langue vietnamienne, langue qui est utilisée par 75 millions de personnes au Vietnam et par environ 3 millions de personnes à l'étranger [Web, 2002]. Le vietnamien est une langue asiatique tonale et monosyllabique comme le mandarin, le cantonnais (un dialecte du chinois utilisé par environ 64 millions personnes sur le monde [Grime, 1992]) ou le thaïlandais. Depuis les années 1990s des systèmes de reconnaissance ont été développés pour le mandarin [Lee et al, 1993][Lyu et al, 1995], puis ces dernières années des recherches ont porté sur l'étude du cantonnais [Lee et al, 1999]. Néanmoins il y a très peu de travaux de recherche sur la reconnaissance de la parole en vietnamien.

Dans le vietnamien, chaque syllabe est prononcée avec un ton lexical. Le ton joue un rôle linguistique. Le vietnamien présente six tons lexicaux. Il y a environ 6800 syllabes avec tons comprenant environ 2400 base-syllabes (syllabes indépendamment du ton). Le vietnamien

étant une langue monosyllabique et bi-syllabique, pour reconnaître les mots il faut réaliser la reconnaissance des syllabes avec ton, c'est-à-dire réaliser la reconnaissance des tons et la reconnaissance de base-syllabes.

Le signal de parole peut être considéré, en première approximation, comme la sortie d'un filtre linéaire variable dans le temps excité par une source en entrée. Ce filtre peut être dérivé du modèle de production de la parole fondé sur la théorie acoustique où la source représente la circulation d'air aux cordes vocales (ou les turbulences de l'air pendant la production des consonnes), et le filtre représente les résonances du conduit vocal. Dans le cas de l'étude des syllabes, nous pouvons dire que le ton est la variation temporelle de la fréquence fondamentale (ou le pitch) de la source d'excitation au niveau des cordes vocales, et que la syllabe sans ton est essentiellement caractérisée par le filtre du conduit vocal. C'est pourquoi, pour la reconnaissance de la base-syllabe, les paramètres spectraux concernant les caractéristiques du filtre seront utilisés, alors que pour la reconnaissance du ton, ce sont les paramètres concernant le pitch qui alimenteront le système de reconnaissance.

Après ce premier chapitre d'introduction nous avons organisé notre mémoire de thèse de la manière suivante :

- Le chapitre 2 est consacré à la présentation des connaissances de base de la phonologie et de la phonétique du vietnamien et à des comparaisons entre le vietnamien et le Mandarin, ainsi que le thaïlandais. Ces connaissances seront utiles pour adapter au mieux notre système de reconnaissance de la parole en vietnamien.
- Le chapitre 3 présente l'état de l'art de la reconnaissance de la parole. Ce chapitre se concentre aussi sur les méthodes de reconnaissance utilisées dans les systèmes conçus pour le Mandarin. Celles-ci peuvent être appliquées au vietnamien, moyennant quelques adaptations.
- Dans le chapitre 4, la caractérisation des tons du vietnamien sera réalisée. Le vietnamien présente trois dialectes principaux qui sont le dialecte du Nord, le dialecte du Sud et le dialecte du Centre. Un corpus est réalisé avec plusieurs locuteurs de ces trois régions. Les analyses des tons sont réalisées sur ce corpus. Les gabarits et les durées des tons sont décrits. Ces résultats serviront pour définir le meilleur vecteur caractéristique pour la reconnaissance des tons du vietnamien.

- La reconnaissance des tons est une tâche importante dans la reconnaissance de la syllabe des langues tonales. Dans le chapitre 5, elle sera réalisée en utilisant une méthode fondée sur un modèle de Markov caché ou HMM. Les résultats des expériences de la reconnaissance des tons dépendants et indépendants du locuteur seront présentés.
- Dans le chapitre 6, un essai de réalisation d'un prototype de reconnaissance de syllabes avec ton du vietnamien en utilisant une méthode fondée sur les HMMs est présenté. Le système de reconnaissance est une combinaison d'un moteur de reconnaissance des tons et d'un moteur de reconnaissance de la base-syllabe. Les résultats de la reconnaissance dépendant du locuteur avec 1168 syllabes avec ton seront discutés.

La conclusion viendra terminer ce mémoire, en décrivant les perspectives possibles de ces travaux.

1.1 Références

Barnett J, Bamberg P, Held M, Huerta J, Manganaro L, Weiss A (1995)
Comparative Performance in Large-Vocabulary Isolated-Word Recognition in Five European Languages
EuroSpeech'95, pp 189-192

Bellik Y (1997)
Multimodal Text Editor Interface Including Speech for the Blind
Speech Communication, vol 23, No. 4, Decmber

Boitet C, Caelen J, Fafiotte, G, Keller, E, Lafourcade, M, Wehrli, E (1998)
Integrating French within C-STAR II
Grenoble,Report and demos of the CLIPS++ group

Grimes B. F (1992)
Ethnologue: Languages of the World
Dallas, Texas: Summer Institute of Linguistics

Hwang, M.Y., X. Huang, and F. Alleva (1993)
Predicting Unseen Triphones with Senones
ICASSP' 93, pp 311-314

Juang B.H (1998)
The Past, Present, and Future of Speech Processing
IEEE Signal Processing Magazine, May 1998, pp 24-48

Lee L-S, Tseng C-Y, Gu H-Y, Liu F-H, Chang C-H, Lin Y-H, Lee Y, Tu S-L, Hsieh S-H et Chen C-H (1993)
Golden Mandarin (I) - A Real Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary
IEEE Trans. on Speech and Audio Processing, vol. 1, no. 2, pp 158-179.

Lee T, Ching P.C (1999)
Cantonese Syllable Recognition Using Neural Networks
IEEE Trans on Speech and Audio Processing, Vol 7, No 4, pp 466-472

Lyu R-Y, Chien L-F Hwang S-H, H H-Y, Yang R-C, Bai B-R, Weng J-C, Yang Y-J, Lin S-W, Chen K-J, Tseng C-Y, Lee L-S (1995)
"Golden Mandarin (III) - A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary"
In Proc of ICASSP, pp 57-60

Schafer R.W (1994)
in "Voice Communication between Humans and Machines"
National Academy Press, Washington D.C 1994

Young S.J, Woodland P.C (1994)
State Clustering in HMM-based Continuous Speech Recognition
Computer Speech and Language, Vol 8, No 4, pp 369-384

Web 2002:

http://www.public.asu.edu/~ickpl/

2

Phonologie et Phonétique vietnamienne

2.1 Introduction

L'origine de la langue vietnamienne est toujours sujet à débat parmi les linguistes. Il est cependant généralement admis qu'elle a des racines communes et fortes avec le *mon-khmer* qui fait partie de la branche austro-asiatique et qui comprend le *mon* parlé en Birmanie et le *kmer* la langue cambodgienne, aussi bien que les *khmu*, *bahnar* et *bru*, d'autres langues parlées par les habitants des îles du nord du Vietnam [Haudricourt 1953, 1954].

La langue vietnamienne est cependant un mélange d'éléments *mon*, *khmer*, *thaï* et *chinois*. Elle a emprunté un bon pourcentage de mots de base aux langues monotoniques *mon* et *khmer*. Des langues *thaï*, elle a adopté certains éléments de grammaire et leur tonalité.

Enfin, bien qu'il est reconnu que le vietnamien n'a pas ses origines en Chine, même si cela est souvent affirmé à tort, le *chinois* a donné au vietnamien l'essentiel de son vocabulaire philosophique, littéraire, technique et gouvernemental, ainsi que son mode d'écriture traditionnel, pendant la période de domination chinoise.

Plus récemment, du fait de l'occupation française en Indochine, certains mots ont été empruntés au français, mais prononcés et orthographiés à la « vietnamienne ». Nous pouvons citer des exemples tels que len (laine), bơ (beurre), bi (bille), cà phê (café), bánh mì (pain (de

mie)), phim (film), mét (mètre), gam (gramme), lít (litre), va li (valise), phó mát (fromage) xúp (soupe), cà rốt (carotte), etc.

Il est suggéré par certains auteurs que le vietnamien était originellement une langue polysyllabique. Les mots polysyllabiques furent simplifiés plus tard par contraction sous l'effet de l'influence des autres langues d'Asie, et plus particulièrement du chinois qui est une langue monosyllabique. Ainsi, des mots qui étaient polysyllabiques vers le 17ème siècle sont maintenant devenus monosyllabiques.

Le vietnamien est donc considéré maintenant comme une langue monosyllabique (ou bisyllabique pour certains mots) possédant six tons, qui donnent à cette langue une impression de langue chantée. Une syllabe peut être répétée avec chacun de ces six tons, ce qui lui donne alors six significations différentes. La syllabe (ou le mot) /ma/ peut prendre six significations comme le montre le tableau **2.1** suivant :

mot vietnamien	traduction en français
ma	fantôme
mà	mais
mã	cheval
mả	tombe
má	joue
mạ	semis

Tableau 2.1 Exemple de différentes significations de la même syllabe /ma/ prononcée avec les 6 tons du vietnamien.

La langue vietnamienne a utilisé les caractères chinois ordinaires (*chu nho*) jusqu'au XIIIe siècle, puis les vietnamiens ont inventé leur propre système d'écriture (*chu nom*) en réunissant des caractères chinois ou en ne les utilisant que pour leur importance phonétique. Les deux systèmes ont cohabité jusqu'au XXe siècle : utilisation du *chu nho* pour l'enseignement et les documents officiels mais la littérature populaire a toujours été écrite avec le *chu nom*.

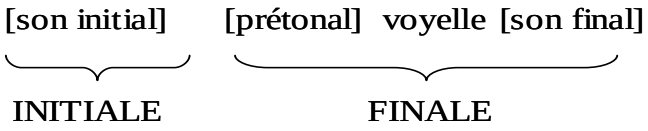
Cependant, le vietnamien moderne s'écrit avec des caractères latins auxquels sont additionnées quelques lettres supplémentaires ou modifiées : le *quoc ngu*. De 1627 à son arrivée au Vietnam où il apprit la langue en moins de six mois, à 1677 où il rédigea le premier dictionnaire vietnamien-latin-portugais, le brillant jésuite avignonnais Alexandre de Rhodes

inventa cette écriture *quoc ngu* dans le but essentiel de faciliter la diffusion de l'Evangile. Cependant, c'est sous la domination française, puis au début du XX^{e} siècle que l'usage de cet alphabet latin s'est généralisé, surtout après la première guerre mondiale, car les vietnamiens étaient persuadés que l'éducation du peuple, grandement facilitée par l'adoption d'un alphabet facile à apprendre, était essentielle au développement de leur pays.

2.2 Structure de la syllabe et phonèmes du vietnamien

La langue vietnamienne est une langue mono ou bisyllabique tonale. Avant de construire un système de reconnaissance de la parole en vietnamien, une connaissance de la phonologie et de la phonétique de la langue est nécessaire. Dans ce chapitre, nous allons présenter la structure de la syllabe puis préciser quelques caractéristiques acoustiques du vietnamien.

Une syllabe peut être composée avec les 6 tons pour créer des mots de signification différente. La structure de la syllabe du vietnamien proposée par [Doan,1977] est la suivante :



Dans cette structure, le son initial, la partie prétonale et le son final sont des options qui peuvent exister ou non dans la syllabe.

2.2.1 Sons initiaux

Il y a 21 consonnes qui peuvent être des sons initiaux, comme le montre le tableau **2.2**. Si nous comptons le cas qui représente l'absence de son initial dans une syllabe et que nous appellerons *phonème zéro,* le système des sons initiaux du vietnamien contient donc 22 phonèmes différents [Nguyen P.P 1992].

phonème	caractère en vietnamien
t’	th
t	t
ʈ	tr
c	ch
k	c,k,q
b	b
d	đ

phonème	caractère en vietnamien
m	m
n	n
ɲ	nh
l	l
ŋ	ng, ngh
f	ph
s	x

phonème	caractère en vietnamien
ʂ	s
χ	kh
h	h
v	v
z	d, gi
ʐ	r
ɣ	g, gh

Tableau 2.2 Les 21 sons initiaux du vietnamien.

Nous pouvons cependant faire deux remarques sur ces phonèmes initiaux :

- Pour les personnes vivant à Hanoi et dans les régions du Nord, les phonèmes c/ʈ, s/ʂ, z/ʐ ne sont pas distincts, c'est-à-dire qu'ils sont prononcés identiquement. Dans les régions Centrale et Sud, ces couples de consonnes sont prononcés bien distinctement.
- Quelques auteurs, [Le, 1948], [Thompson, 1965], rajoutent à ce système des consonnes initiales une consonne /ʔ/ qui est considérée comme une consonne de coup de glotte. Cette 22ème consonne ne correspond à aucun caractère du système orthographique *quoc ngu*. Avec cet ajout, ces auteurs considèrent que la partie initiale de la structure de la syllabe est toujours existante dans toutes les syllabes. Par exemple, dans ce cas le mot "an" pour transcription phonétique /ʔan/. Cette stratégie étant toujours débattue, nous avons décidé de conserver le système à 21 consonnes initiales parce qu’il correspond d’une façon simple à la description faite par le système orthographique *quoc ngu*.

Pour le lecteur qui voudrait se faire une idée des différents sons de la langue vietnamienne, nous pouvons citer le site Internet Vietnamese Language and Culture [http://www.seasite.niu.edu/vietnamese/VNMainPage/vietsite/vietsite.htm]. Le lecteur pourra ouvrir la page *Guide_to_prononciation* > *Consonants* : des exemples sonores sont téléchargeables pour une écoute.

Point d'articulation / **Modes d'articulation**			Labiales	Apicales		Palatales	Dorsales	Glottale
				dentales	rétro-flexes			
Occlusives	aspirée sourde			t'				
	non aspirées	sourdes	p	t	ʈ	c	k	ʔ
		sonores	b	d				
	nasales		m	n		ɲ	ŋ	
Fricatives	sourdes		f	s	ʂ		χ	h
	sonores		v	z	ʐ		ɣ	
	latéral			l				

Tableau 2.3 Lieu d'articulation des consonnes vietnamiennes (après Doan, 1977)

L'occlusive labiale sourde /p/ n'existe pas en vietnamien standard mais est prononcée avec les mots d'origine étrangère

2.2.2 Prétonal

Le système des sons prétonaux comprend 2 phonèmes avec un phonème *zéro* /ø/ représentant l'absence de la partie prétonale. La semi-voyelle labiale /-u̯-/ dans le vietnamien est considérée comme prétonale (tableau **2.4**).

semi-voyelle	caractère
-u̯-	**u, o**

Tableau 2.4 Prétonal du vietnamien

La distribution du système prétonal qui suit les sons initiaux est montrée par le tableau **2.5** [Doan 1977] :

	ø	t'	t	ʈ	c	k	b	d	m	n	ɲ	ŋ	f	s	ʂ	χ	h	v	z	ʐ	ɣ	l
ø	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
-u̯-	+	+	+	+	+	+	-	+	-	-	+	+	-	+	+	+	+	-	+	-	+	+

Tableau 2.5 Combinaisons possibles entre les sons prétonaux et les sons initiaux.

Le signe "+"signifie que la combinaison est possible, le signe "-" signifie que celle est impossible

Les remarques suivantes peuvent apporter quelques éclairages particuliers :

- le phonème *prétonale zéro* peut apparaître derrière tous les sons initiaux, c'est-à-dire que la semi-voyelle /-u̯-/ peut ne pas exister dans une syllabe.
- cependant, la semi-voyelle /-u̯-/ n'apparaît jamais derrière les consonnes /b, m, f, v, n, ʐ/ (sauf dans les mots étrangers qui ont été transcris comme par exemple "buýt" qui veut dire "bus" en français).
- l'association de la semi-voyelle /-u̯-/ avec la consonne initiale /n/ apparaît seulement pour le mot d'origine chinoise "noãn" (ovule en français).
- la consonne /ʐ/ n'apparaît devant la semie-voyelle /-u-/ uniquement dans le mot "roạt" (bruit de faucille coupant les céréales).
- la semi-voyelle /-u̯-/ peut apparaître après toutes les consonnes initiales restantes.

2.2.3 Voyelles

Le vietnamien possède 16 voyelles que nous pouvons décomposer en 13 voyelles (tableau ***2.6***) et 3 diphtongues (tableau **2.8**). Les voyelles longues peuvent apparaître individuellement pour créer les mots. Cependant, la langue vietnamienne présente 4 voyelles courtes qui ne peuvent pas apparaître individuellement et qui doivent être combinées avec les sons de la partie finale de la syllabe. Le tableau ***2.6*** montre la classification entre les voyelles longues et les voyelles courtes, le tableau **2.7** donne une idée de l'articulation de ces voyelles.

voyelle		caractère
longue	i	y, i
	ɯ	ư
	u	u
	e	ê
	ɤ	ơ
	o	ô
	ɛ	e
	a	a
	ɔ	o
courte	ɤ̆	â
	ɛ̆	a (anh, ach)
	ă	ă, a (au, ay)
	ɔ̆	o (ong, oc)

Tableau 2.6 Voyelles du vietnamien.

	Position de la langue	antérieures	postérieures	postérieures
	Position des lèvres	*étirées*	*étirées*	*arrondies*
Ouverture de la bouche	fermée	i	ɯ	u
	demi-fermée	e	ɤ, ɤ̆	o
	demi-ouverte	ɛ, ɛ̆		ɔ, ɔ̆
	ouverte		a, ă	

Tableau 2.7 Lieu d'articulation des voyelles vietnamiennes

diphtongue	caractère
i‿e	**iê, ia, yê, ya**
ɯ‿ɤ	**ươ, ưa**
u‿o	**uô, ua**

Tableau 2.8 Diphtongues du vietnamien.

	i	ɯ	u	e	ɤ	o	ɛ	a	ɔ	ɤ̆	ɛ̆	ă	ɔ̆	i‿e	ɯ‿ɤ	u‿o
ø	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
-u̯-	+	-	-	+	+	-	+	+	-	+	+	+	+	+	-	-

Tableau 2.9 Ensemble des combinaisons possibles entre la partie prétonale et les voyelles.

Remarques :

- derrière la semi-voyelle /-u̯-/, il n'existe jamais les voyelles /u, o, ɔ, ɯ/, ni les 2 diphtongues /ɯ‿ɤ, u‿o/ ;
- de même, pour écouter des exemples sonores le lecteur pourra se reporter à la page *Guide_to_prononciation* > *Vowels* du site Internet Vietnamese Language and Culture [http://www.seasite.niu.edu/vietnamese/VNMainPage/vietsite/vietsite.htm].

2.2.4 Sons finaux

Six consonnes et deux semi-voyelles, appelées *sons finaux*, peuvent être présentes dans la partie finale de la syllabe vietnamienne, comme le décrit le tableau **2.10**

sons finaux		caractère
consonne	p	p
	t	t
	k	ch, c
	m	m
	n	n
	ŋ	nh, ng
semi-voyelle	u̯	o, u
	i̯	y, i

Tableau 2.10 Les 8 sons finaux (consonnes et semi-voyelles)

Le système des sons finaux de la langue vietnamienne contient donc 8 phonèmes, auxquels nous pouvons rajouter un phonème *zéro* qui représente la présence d'aucun des 8 sons finaux dans la syllabe. A partir de l'ensemble des voyelles vietnamiennes et des diphtongues, nous pouvons présenter dans le tableau **2.11**, la distribution des combinaisons possibles avec les 8 sons finaux, situés après les voyelles ou diphtongues :

	i	ɯ	u	e	ɤ	o	ɛ	a	ɔ	ɤ̆	ɛ̆	ă	ŏ	i͡e	ɯ͡ɤ	u͡o
ø	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+
p	+	-	+	+	+	+	+	+	+	+	-	+	-	+	+	?
t	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+
k	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+
m	+	-	+	+	+	+	+	+	+	+	-	+	-	+	+	+
n	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+
ŋ	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	+
u̯	+	+	-	+	+	-	+	+	-	+	-	+	-	+	+	-
i̯	-	+	+	-	-	+	-	+	+	+	-	+	-	-	+	+

Tableau 2.11 Ensemble des combinaisons possibles entre les voyelles et les sons finaux.

Nous pouvons constater que :

- les voyelles courtes sont toujours combinées avec des sons finaux, elles ne peuvent pas apparaître individuellement et constituer elles-mêmes la partie finale de la syllabe ;
- /ɛ̆/ et /ɔ̆/ apparaissent seulement devant les consonnes / k/ et /ŋ/.

2.2.5 Syllabes complètes

A partir des descriptions des sons des différentes parties des syllabes de la langue vietnamienne, faites dans ces paragraphes 2.2, 2.3 et 2.4 précédents, il est alors possible de réaliser la combinaison totale *prétonal*, *voyelle* et *son final* pour créer **la partie finale** complète de la syllabe. Nous obtenons ainsi 155 parties finales (on trouve parfois le chiffre 153 dans la littérature. Les 2 parties finales "contestées" sont accompagnées d'un astérisque dans le tableau **2.12**) parties finales différentes qui sont décrites dans le tableau **2.12**.

Ces parties finales peuvent être combinées aussi avec les 22 parties initiales décrites au paragraphe 2.1 pour créer les mots monosyllabiques du vietnamien.

Dans la littérature sur la langue vietnamienne, on considère habituellement deux types de syllabes du vietnamien, les syllabes ouvertes et les syllabes fermées :

- une syllabe fermée présente dans sa partie finale un son final constitué de l'une de 3 consonnes telles que /p/, /t/ et /k/ ;
- les syllabes restantes sont appelées les syllabes ouvertes.

	ø	p	t	k	m	n	ŋ	u̯	i̯
i	i	ip	it	ik	im	in	iŋ	iu̯	
e	e	ep	et	ek	em	en	eŋ	eu̯	
ɛ	ɛ	ɛp	ɛt	ɛk	ɛm	ɛn	ɛŋ	ɛu̯	
ɛ̆				ɛ̆k			ɛ̆ŋ		
ɯ	ɯ		ɯt	ɯk		ɯn	ɯŋ	ɯu̯	ɯi̯
ɤ	ɤ	ɤp	ɤt		ɤm	ɤn			ɤi̯
ɤ̆		ɤ̆p	ɤ̆t	ɤ̆k	ɤ̆m	ɤ̆n	ɤ̆ŋ	ɤ̆u̯	ɤ̆i̯
a	a	ap	at	ak	am	an	aŋ	au̯	ai̯
ă		ăp	ăt	ăk	ăm	ăn	ăŋ	ău̯	ăi̯
u	u	up	ut	uk	um	un	uŋ		ui̯
o	o	op	ot	ok	om	on	oŋ		ui̯
ɔ	ɔ	ɔp	ɔt	ɔk˚	ɔm	ɔn	ɔŋ˚		ɔi̯
ŏ				ŏk			ŏŋ		
i‿e	i‿e	i‿ep	i‿et	i‿ek	i‿em	i‿en	i‿eŋ	i‿eu̯	
ɯ‿ɤ	ɯ‿ɤ	ɯ‿ɤp	ɯ‿ɤt	ɯ‿ɤk	ɯ‿ɤm	ɯ‿ɤn	ɯ‿ɤŋ	ɯ‿ɤu̯	ɯ‿ɤi̯
u‿o	u‿o		u‿ot	u‿ok	u‿om	u‿on	u‿oŋ		u‿oi̯
u̯i	u̯i		u̯it	u̯ik			u̯iŋ	u̯iu̯	
u̯e	u̯e		u̯et	u̯ek		u̯en	u̯eŋ		
u̯ɛ	u̯ɛ		u̯ɛt			u̯ɛn			
u̯ɛ̆				u̯ɛ̆k			u̯ɛ̆ŋ		
u̯ɤ	u̯ɤ								
u̯ɤ̆			u̯ɤ̆t			u̯ɤ̆n	u̯ɤ̆ŋ		u̯ɤ̆i̯
u̯a	u̯a	u̯ap	u̯at	u̯ak	u̯am	u̯an	u̯aŋ	u̯au̯	u̯ai̯
u̯ă	u̯ă	u̯ăp	u̯ăt	u̯ăk	u̯ăm	u̯ăn	u̯ăŋ		u̯ăi̯
u̯i‿e	u̯i‿e		u̯i‿et			u̯i‿en	u̯i‿eŋ		

Tableau 2.12 Ensemble des 155 parties finales des syllabes du vietnamien

2.2.6 Les tons de la langue vietnamienne

Il existe en vietnamien six tons différents dont cinq sont indiqués chacun par un signe. Le premier ton *bằng (plat ou égal)* est sans signe distinctif dans l'alphabet et correspond au ton1 dit aussi *không dấu* (c'est-à-dire *sans signe*) ou *ngang (horizontal)* :

- le ton *bằng*, dit *égal*, reste au même niveau prosodique pendant la prononciation de la syllabe ;
- le ton *huyền*, dit *descendant*, se produit à un registre plus bas que le ton *bằng;*
- le ton *ngã*, dit *brisé* (ou *retombant*), voit son origine vers le ton *bằng*, monte encore, mais il est interrompu comme par un coup de glotte puis relâché vers le haut ;
- le ton *hỏi*,, dit *interrogatif*, peut être considéré comme un passage hésitant du ton *huyền* vers une intonation plus haute ;
- le ton *sắc*, dit *aigu*, part de la hauteur du ton et monte d'un trait rapide ;
- le ton nặng, dit *grave*, produit un son étranglé dans la gorge (les vibrations des cordes vocales sont interrompues brusquement).

Ton	Description	Tiếng Việt	Signe
ton1	ton plat	không dấu	
ton2	ton descendant	huyền	`
ton3	ton brisé	ngã	~
ton4	ton interrogatif	hỏi	?
ton5	ton montant	sắc	‘
ton6	ton grave	nặng	.

Tableau 2.13 Description succincte des six tons vietnamiens, avec leurs noms et signes associés.

A partir de la figure 2.1 représentant très schématiquement l'évolution temporelle des 6 tons, ainsi que leur hauteur relative, nous pouvons classer les 6 tons dans deux catégories : tons du registre haut et tons du registre bas.

		monotone	mélodique	glottal
registre	haut	ton1	ton5	ton3
	bas	ton2	ton4	ton6

Tableau 2.14 Classification des six tons en 2 registres et 3 catégories tonales.

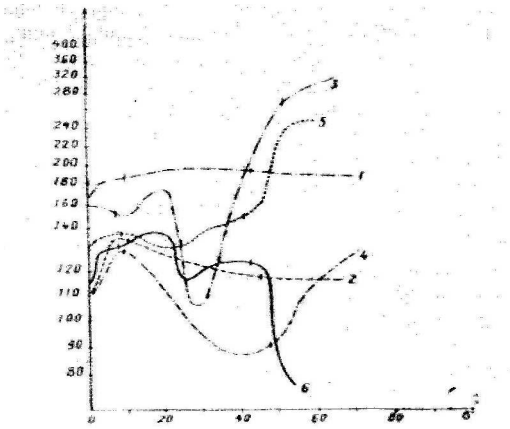


Figure 2.1 Evolution temporelle des tons du vietnamien (d'après [Andreev,1957])

En général, chaque syllabe peut être combinée avec les six tons pour créer les six mots différents. Cependant, les syllabes fermées ne peuvent se réaliser en combinaison des deux tons : ton5 et ton6.

2.3 Différences acoustiques entre les dialectes des régions Nord, Centrale et Sud du Vietnam

Carte 1 : Les 3 régions du Vietnam. Nord (Hanoi), Centre (Da Nang) et Sud (Ho Chi Minh Ville)

La langue vietnamienne est habituellement divisée en trois dialectes correspondant aux 3 régions Nord, Centre et Sud (carte 1). Des différences sensibles existent entre les 3 prononciations :

- pour les personnes vivant à Hanoi et dans les régions du Nord, les phonèmes c/ʈ, s/ʂ, z/ʐ ne sont pas distinctes, c'est-à-dire qu'ils sont prononcés identiquement. Par exemple les mots "xôi" /soi̯/ (riz gluant cuit à la vapeur) et "sôi" /ʂoi̯/ (bouillir) présentent exactement la même acoustique. Les phonèmes /ʂ, ʐ/ sont prononcés comme les phonèmes /s, z/ qui sont les consonnes apicales-dentales, le phonème /ʈ/ est prononcé comme le phonème /c/.
- dans les régions du Centre et du Sud, ces couples de phonèmes sont prononcés bien distinctement et trois phonèmes /ʈ, ʂ, ʐ/ sont prononcés comme des rétroflexes ;

- dans les écoles primaires, il est enseigné la prononciation d'un son initial correspondant au caractère "r" sous la forme d'un son vibrant proche du "r" roulé. Mais cette prononciation est en fait peu utilisée car elle est différente de la prononciation populaire [Doan, 1977]. La prononciation de ce son vibrant existe seulement dans quelques petites régions du Centre. En pratique, la prononciation du caractère "r" correspond à la prononciation de la phonème /ʐ/. La figure **2.2** donne une comparaison de la prononciation des /ʂ, ʐ/ des personnes du Centre et du Sud avec la personne du Nord.

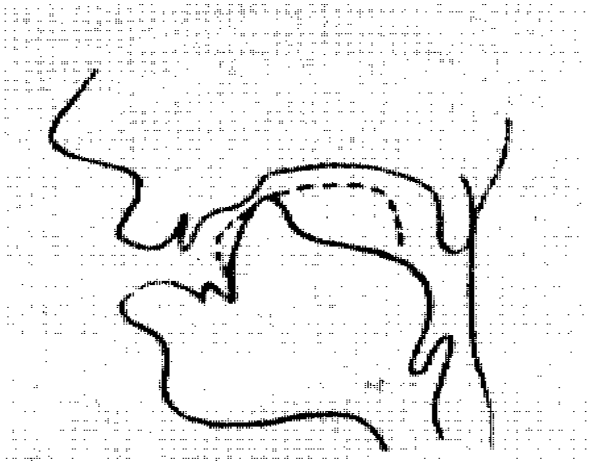


Figure 2.2 Comparaison de la prononciation de /ʂ/ des personnes Central et Sud avec la personne Nord (tirets). (d'après [Doan, 1977])

Au niveau de la variation temporelle des tons, une seule différence importante existe entre les régions : le ton3 et le ton4 sont prononcés identiquement par les personnes originaires des régions Sud et Centre, alors que ces deux tons sont prononcés bien distinctement par les vietnamiens du Nord. Nous vérifierons cette constatation dans l'étude de notre corpus de mots isolés, enregistrés avec des locuteurs des trois régions.

2.4 Quelques particularités principales des langues Mandarin et Thaïlandais

Dans ce paragraphe, nous souhaitons présenter une rapide comparaison entre la langue vietnamienne et les langues chinoises et thaï. La connaissance des ressemblances entre elles nous permet d'appliquer efficacement les techniques de reconnaissance existantes de la langue chinoise ou du thaï au vietnamien.

2.4.1 Mandarin

Le Mandarin est la langue officielle chinoise. Cette langue est écrite en caractères chinois traditionnels, composés d'un ensemble important de logographes. Comme le vietnamien, le Mandarin est une langue monosyllabique à tons. Tous les caractères chinois sont monosyllabiques avec un nombre de monosyllabes possibles d'environ 1345 syllabes. En général, chaque syllabe est prononcée avec l'un des tons. A la différence du vietnamien, le Mandarin n'a que quatre tons lexicaux et un ton neutre. Le ton neutre existe seulement dans la syllabe terminale d'un mot multi-syllabique et il n'a pas un contour fixe. Les 1345 syllabes avec tons possibles fondées sur 408 syllabes indépendamment du tons [Lee, 1997].

La structure de la syllabe sans ton est similaire à celle du vietnamien, avec une partie initiale et une partie finale [Yang et all., 1988] :

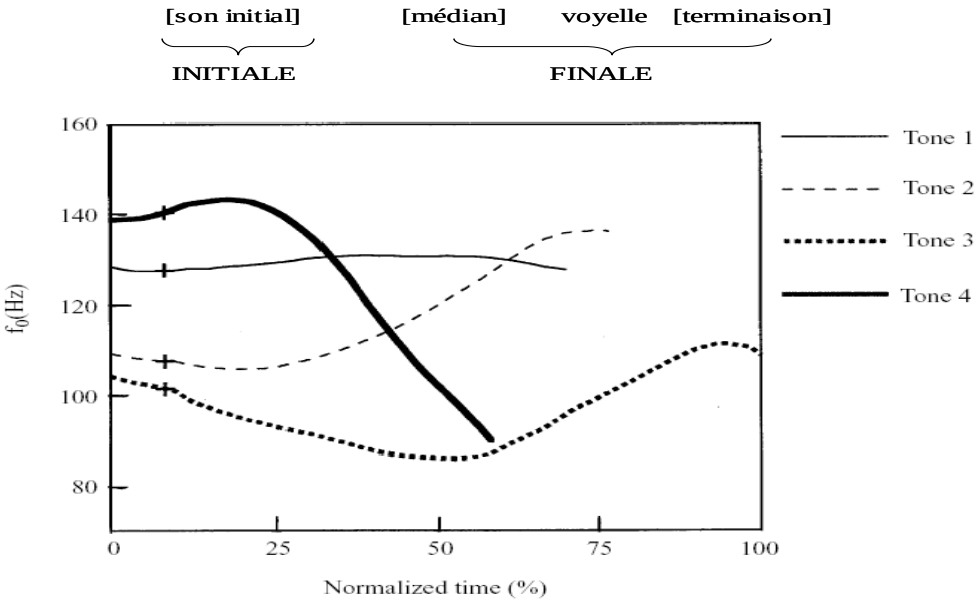


Figure 2.3 Contours de quatre tons du Mandarin pour la syllabe [ma] (après [Xu, 1997])

Seule la voyelle ou la diphtongue de la partie finale est obligatoire, alors que les autres composantes entre parenthèses sont optionnelles. Le Mandarin compte 22 parties initiales (comptées avec un phonème nul) et 38 parties finales [Lee, 1997].

Les contours des tons du Mandarin sont présentés en figure **2.3**.

2.4.2 Langue thaïlandaise

Le Thaï est une langue dans la famille de la langue Tai. C'est la langue officielle en Thaïlande. Le Thaï utilise un script qui est alphabétique. Le Thaï est une langue tonale et principalement monosyllabique, par ailleurs il existe aussi des mots polysyllabiques qui viennent du Khmer, du Pali ou du Sanskrit [Karoonboonyanan, 1999].

La structure de la syllabe thaï est présentée par Tungthangthum (1998) de la façon suivante :

[consonne] [consonne] voyelle [consonne]

Cette structure est différente de celle de la syllabe vietnamienne : la syllabe thaï contient un agglomérat de consonnes, une voyelle en position médiane et une consonne finale. Les consonnes peuvent être présentes ou non dans la syllabe. Le Thaï présente 20 consonnes, 21 voyelles parmi lesquelles il y a 9 voyelles courtes, 9 voyelles longues et 3 diphtongues [http://thaiarc.tu.ac.th/host/thaiarc/thai].

Le Thaï est prononcé avec cinq tons différents dont les allures des variations temporelles sont présentées par la figure **2.4**.

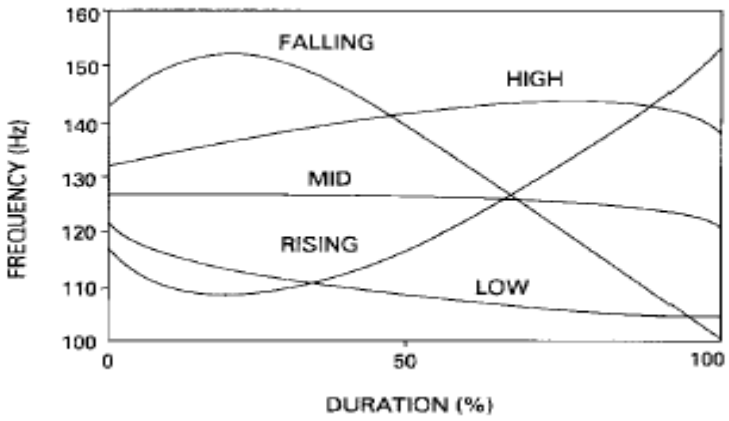


Figure 2.4 Contours temporels des cinq tons de la langue Thaï (après [Potisukl, 1997])

2.5 Conclusion

Dans ce chapitre nous avons présenté les caractéristiques principales de la phonologie et de la phonétique de la langue vietnamienne. C'est une langue tonale, monosyllabique ou bisyllabique présentant six tons (vietnamien standard). La structure de la syllabe vietnamienne comprend une partie initiale et une partie finale. Nous avons présenté les règles de combinaisons entre les consonnes et les voyelles pour créer une syllabe.

Nous avons résumé quelques caractéristiques du Mandarin (la langue officielle en Chine) et Thaï (la langue officielle du Thaïlande). Grâce à ces ressemblances, les techniques de reconnaissance appliquées au Mandarin et au Thaï pourront probablement être appliquées à la reconnaissance du vietnamien.

2.6 Références

Andreev N.D, Gordina M. V
"Système des tons vietnamiens" (en russe)
по эсперименталъым данным, Весник, No 8

Doan T.T (1977)
"Ngữ âm tiếng Việt"
Nha Xuat Ban Editions, 1977.

Haudricourt A.G (1953)
La place du vietnamien dans les langues austroasiatiques
Bulletin de la Société de Linguistique de Paris, 1953, 49, 1

Haudricourt A.G (1954)
De l'origine des tons en vietnamiens
Journal Asiatique, 1954, 242, 1

Karoonboonyanan T (1999)
Standardization and Implementations of Thai Language
presented at the Seminar on Enhancement of the International Standardization Activities in Asia Pacific Region **(AHTS-1)** held on at CICC, Japan, in March 1999.

Le V.L (1948)
Le parler vietnamien
Paris, 1948

Lee L-S (1997)
Voice Dictation of Mandarin Chinese
IEEE Signal Processing Magazine, July 1997

Potisuk S, Harper M.P, Gandour J (1997)
Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method
IEEE Trans on Speech and Audio Processing, vol 7, No 1, pp 95-102

Thompson L.C (1965)
A Vietnamese Grammar
University of Washington Press, Seattle, 1965

Tungthangthum A (1998)
Tone Recognition for Thai
Circuits and Systems, IEEE APCCAS 1998, Asia-Pacific Conference, p. 157-160.

Web, 2000
http://www.seasite.niu.edu/vietnamese/VNMainPage/vietsite/vietsite.htm

Web, ThaiARC
http://thaiarc.tu.ac.th/host/thaiarc/thai

Xu Y (1997)
Contextual tonal variations in Mandarin
Journal of Phonetics, vol 25, pp 61-83

Yang W. J, Lee J. C, Chang Y. C et Wang H. C (1988).
Hidden Markov Model for Mandarin Lexical Tone Recognition
IEEE Trans. ASSP, vol 36, no 7, pp 988-992

3

Reconnaissance de la parole

3.1 Introduction

Les études scientifiques sur la reconnaissance de la parole ont commencé au début des années 1950, mais ont donné leurs premiers résultats tangibles dans les années 70 avec le projet ARPA (Advanced Research Projects Agency [Newell et al., 1973 ; Klatt, 1977]. Leur principal but a été, et est toujours, de concevoir un système de reconnaissance de la parole robuste et efficace, pour obtenir un outil s'intégrant *naturellement* dans la communication entre l'homme et la machine et facilitant son interaction. Pour le commun des hommes, les robots et des machines qui communiquent en langage naturel avec les acteurs des films de science-fiction, est l'image immédiate qui s'impose. Cependant, sans tomber dans cet extrême, de nombreuses applications d'un tel outil de reconnaissance automatique de la parole (RAP) peuvent dès maintenant être envisagées.

Ces applications peuvent être regroupées en plusieurs catégories :

- *Dictée vocale* : le marché de l'informatique propose maintenant des logiciels sous Windows, capables, pour une modique somme à l'achat, de proposer des fonctionnalités de création de documents avec une seule interface parole ;
- *Commande de machines et contrôle de processus* : dans l'industrie, il n'est pas toujours possible de piloter les machines avec les moyens habituels (série de boutons, clic avec la souris, choix dans un menu, etc.), d'autant plus que les machines deviennent de plus en

plus sophistiquées et que leur usage (ou leur processus de commande) devient de plus en plus complexe. La parole, grâce à un système de commande orale, peut s'avérer un mode de commande rapide et concis, entrouvrant un espace de liberté supplémentaire à l'utilisateur ;

- *Reconnaissance automatique de la parole dans les télécoms:* depuis les années 90, nous sommes entrés dans un monde où l'usage des nouvelles technologies de l'information se développe de façon exponentielle. Le téléphone portable, devient un objet courant et traduit une volonté de nomadisme et de liberté pour l'usager. L'une des conséquences est que les technologies vocales sont maintenant sorties des laboratoires de recherche et permettent aux "providers" de services de les mettre à disposition de leurs clients pour faciliter l'usage des moyens de télécommunication : la composition automatique des numéros de téléphone, les serveurs vocaux pour la réservation des billets de transports ou pour la consultation de services bancaires, sont des exemples devenus maintenant courants.
- *Traduction automatique:* ce dernier type d'applications conjugue les nouvelles technologies de la traduction automatique de la langue avec les technologies de la reconnaissance automatique et de la synthèse de la parole. Même si ces systèmes ainsi conçus ne permettent pas de la traduction exacte au mot près, ils permettent d'aider au dialogue entre deux personnes de langue maternelle différente, en proposant une traduction approchée des thèmes et des concepts évoqués dans le dialogue. Nous pouvons citer le projet international C-Star (Consortium for Speech Translation Advanced Resarch) auquel participe le laboratoire CLIPS [Boitet et al, 1998].

3.2 Caractérisation de la RAP

La reconnaissance automatique de la parole (RAP) est un processus qui convertit le signal acoustique de parole en un ensemble de mots. Les systèmes de RAP peuvent être caractérisés par plusieurs paramètres.

3.2.1 Mode et style de prononciation

Il existe trois modes de prononciation distincts :

- *mots isolés :* chaque mot est prononcé isolément ; une pause de durée importante sépare les mots ;
- *mots connectés :* le système reconnaît des séquences de quelques mots sans pause volontaire entre eux (exemple : reconnaissance de chiffres connectés ou de nombres quelconques, etc.) ;
- *parole continue :* les mots sont prononcés naturellement sous forme de phrases aux énoncés plus ou moins longs.

Au mode de présentation, peut être associé un style : la parole peut être le résultat de la lecture d'un document, cas pour lequel les contraintes orthographiques, grammaticales et linguistiques sont le plus souvent assez fortes, mais la parole peut être plus simplement de la parole spontanée (dialogue par exemple) où les contraintes de structure sont souvent estompées.

3.2.2 Mode d'utilisateurs

Le système de reconnaissance de la parole peut se distinguer suivant deux types d'utilisation différents :

- *dépendant du locuteur*: le système est adapté à la voix d'un locuteur particulier.
- *indépendant du locuteur* : le système n'est pas adapté à la voix d'un locuteur particulier et peut être utilisé par un locuteur quelconque.

3.2.3 Taille du vocabulaire

Suivant sur la taille de vocabulaire, les systèmes de RAP sont classifiés en trois types [Rabiner et Juang ,1993]:

- système de RAP avec un petit vocabulaire lequel contient de 10 à 100 mots ;
- système de RAP avec un moyen vocabulaire lequel contient de 100 à 1000 mots ;
- système de RAP avec un grand vocabulaire lequel contient plus de 1000 mots ;

3.2.4 Modèle de langage

Dans un système de reconnaissance de la parole, le modèle de langage est utilisé pour limiter le nombre des combinaisons entre les mots d'une phrase. Un modèle de langage peut être de type N-grammes (tout mot peut être suivi par une séquence de N-1 mots autres avec une probabilité fixée) ou fondé sur des grammaires.

3.2.5 Environnement

Les changements de conditions environnementales (bruit, changement de microphone, acoustique de la salle, etc.) peuvent influer sur le taux de reconnaissance du système. La plupart des systèmes vont fonctionner correctement dans un environnement aux caractéristiques acoustiques et sonores proches de l'environnement dans lequel s'est fait l'entraînement mais les performances vont se dégrader notablement si les conditions environnementales sont très différentes.

De nombreuses équipes de recherche étudient des moyens de rendre les systèmes de reconnaissance insensibles, c'est-à-dire robustes, aux changements des conditions environnementales, soit en réalisant des moteurs de reconnaissance capables de s'adapter eux-mêmes aux nouvelles conditions avec, par exemple, des techniques d'apprentissage adaptative [Gales, 2001], des techniques de modèles évolutionnaires [Spalanzani, 1999], soit en faisant précéder les moteurs de reconnaissance par des modules de traitement du signal capables d'améliorer le signal de parole à l'entrée du système de reconnaissance [Nguyen Thi, 1993].

Ces dernières techniques étudient des modules de réhaussement de la parole dans le bruit, ou bien des modules d'annulation d'écho des salles, ou encore du filtrage adaptatif pour éliminer les bruits. Nous présentons dans notre Annexe, un résumé du travail que nous avons mené au sujet de la robustesse, au sein de l'équipe GEOD.

3.3 Structure d'un système de RAP

Un système de reconnaissance de la parole contient un module d'extraction des paramètres acoustiques, un module de décodage et trois sources d'informations : les modèles acoustiques, le lexique et le modèle de langage (figure **3.1**)

- *le module d'extraction des paramètres acoustiques* permet de convertir le signal de parole sous la forme de vecteurs acoustiques qui représentent les informations nécessaires et suffisantes pour les processus de reconnaissance ;
- *les modèles acoustiques* représentent les unités phonétiques qui peuvent être un mot, une syllabe, un phonème ;
- *le lexique* est utilisé pour créer les mots à partir des modèles acoustiques. Pour un système utilisant un petit dictionnaire, le lexique fait habituellement la correspondance entre un modèle acoustique et un mot. Pour un grand vocabulaire, le modèle acoustique étant habituellement un phonème, le lexique impose alors la combinaison des phonèmes pour créer un mot [Lamel, et al, 1996] ;
- *le modèle de langage* contient les informations qui indiquent comment connecter les mots ensemble dans un arbre des mots possibles. Le modèle de langage peut contenir les syntaxes grammaticales ou des modèles de langage stochastiques comme les N-grammes [Boite et al, 2000].
- à partir des vecteurs acoustiques et des différents modèles, *le décodage* est effectué pour sortir la phrase la plus probable.

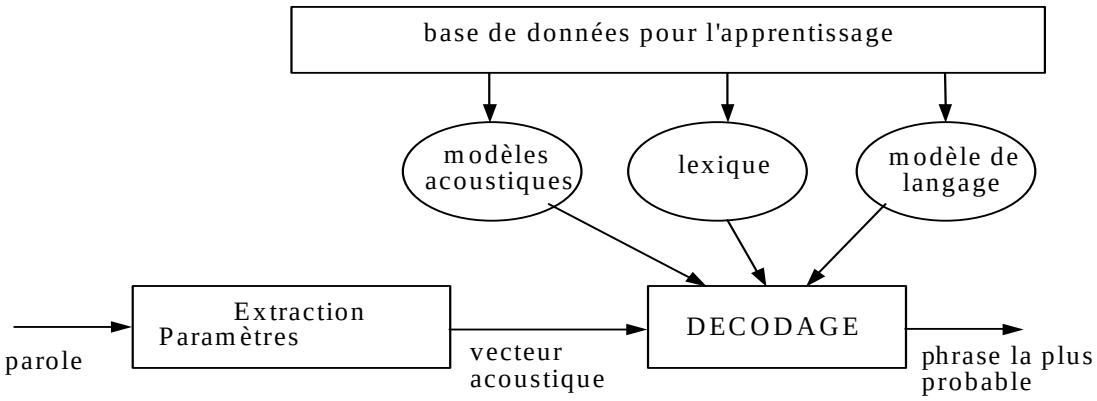


Figure 3.1 Schéma général d'un système de reconnaissance de la parole

3.3.1 Extraction des paramètres acoustiques

Le signal de parole est complexe et redondant. Il possède une grande variabilité. Pour que le système de reconnaissance de la parole fonctionne efficacement, les informations caractéristiques et invariantes doivent être extraites du signal de parole. Cette procédure consiste à associer au signal de parole une série de vecteurs de paramètres généralement acoustiques ou spectraux. La figure **3.2** propose le schéma d'un module d'extraction de ces paramètres avec :

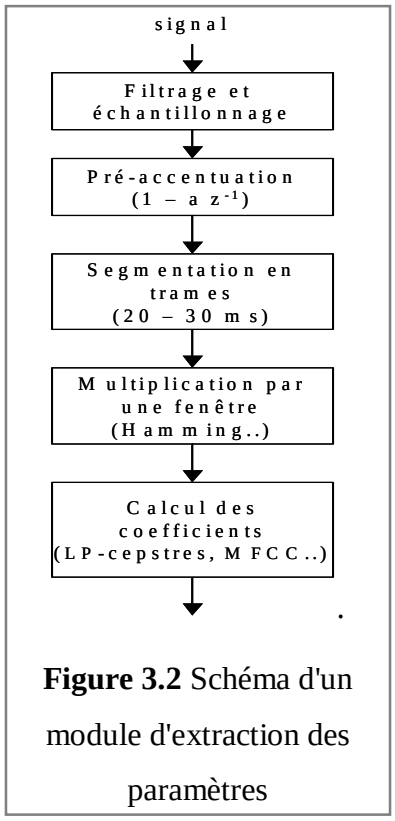


Figure 3.2 Schéma d'un module d'extraction des paramètres

- une phase **de filtrage, d'échantillonnage et de quantification** qui correspond à la première étape de numérisation du signal ;
- un module de **pré-accentuation :** il y a deux explications pour l'utilisation du module de pré-accentuation [Picone, 1993]. Pour la première, la partie voisée du signal de la parole présente une accentuation spectrale approximative de -20 dB par décade. Le filtre de pré-accentuation permet de compenser cette accentuation avant d'analyser le spectre, ce qui améliore cette analyse. La deuxième considère que l'audition est plus sensible dans la région du spectre autour des 1 kHz. Le filtre pré-accentuation va donc amplifier cette région centrale du spectre.
- une troisième opération **de segmentation** en trames permet de découper le flot de parole continue en trames pendant lesquelles le signal est supposé quasi-stationnaire ; chaque trame a habituellement une durée identique d'environ 20 à 30 ms ;
- pour réduire les effets de bord produits par la segmentation, les trames sont alors multipliées par **une fenêtre** (de Hamming le plus souvent) ;
- enfin, le **calcul des coefficients** est effectué : il existe plusieurs types de coefficients, LPC, PLP, MFCC, etc.

3.3.1.1 Les coefficients de prédiction linéaire (LPCs)

L'analyse LPC [Makhoul, 1975] est fondée sur le modèle de production de la parole, qui considère que l'appareil de production de la parole (cordes vocales + conduit vocal complet) est constitué d'une source (source pseudo-périodique ou source de bruit) et d'un filtre se comportant comme un résonateur (conduit vocal). La figure **3.3** schématise ce modèle simplifié de la production de parole. Le signal de la parole peut être ainsi modélisé comme étant le signal en sortie d'un filtre H(z) dont la source d'excitation à l'entrée du filtre u(t) est soit une source de série d'impulsions quasi-périodiques ou soit une source de bruit aléatoire : les sons voisés (voyelles) sont créés par la source d'impulsions quasi-périodiques et les sons non-voisés sont le résultat d'une turbulence ou d'une explosion dans le conduit vocal et sont caractérisés par une source de bruit aléatoire.

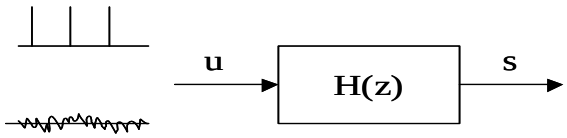


Figure 3.3 Modèle de production de la parole

L'analyse LPC repose sur l'hypothèse que le filtre est un filtre tout-pôles :

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \qquad (3.1)$$

où G est le coefficient de gain, $\{a_k\}$sont les coefficients LPC et p est l'ordre du filtre.

Avec cette hypothèse, le signal de la parole peut être considéré comme un signal auto-régressif :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (3.2)$$

Les coefficients a_k et le gain G sont calculés grâce à des méthodes fondées sur le calcul de la matrice de covariance [Atal & Hanauer, 1971] ou grâce à des méthodes fondées sur le calcul de la matrice d'autocorrélation [Itakura & Saito, 1968].

Les coefficients cepstraux, c_m, peuvent être calculés fondés sur les coefficients LPCs [Rabiner, 1993]:

$$\begin{aligned} c_0 &= \ln G^2 \\ c_m &= a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \qquad 1 \le m \le p \\ c_m &= \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \qquad m \succ p \end{aligned} \tag{3.3}$$

3.3.1.2 Les coefficients PLP (Perceptual Linear Predictive)

PLP est une technique d'analyse de la parole [Hermansky, 1990] fondée sur la modélisation du spectre par un modèle tout-pôle suivant un principe identique à la technique de prédiction linéaire (LP). Cependant, la différence réside dans le fait que les paramètres d'un filtre auto-régressif tout pôle sont estimés en modélisant au mieux le spectre auditif. Ceci est fondé sur trois effets auditifs : sélectivité spectrale de bande critique, courbe d'intensité égale et loi de puissance.

La figure **3.4** représente le processus de calcul des coefficients PLP. Pour obtenir un spectre auditif, la courbe de masquage $\Psi(\Omega)$ est tout d'abord utilisée

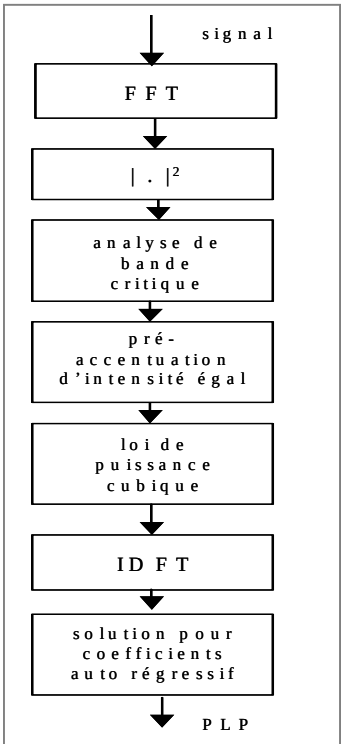


Figure 3.4 Calcul des coefficients PLP

$$\Psi(\Omega) = \begin{cases} 0 & si\ \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & si\ -1.3 \le \Omega \le -0.5 \\ 1 & si\ -0.5 \le \Omega \le 0.5 \\ 10^{-1.0(\Omega-0.5)} & si\ 0.5 \le \Omega \le 2.5 \\ 0 & si\ \Omega > 2.5 \end{cases} \tag{3.4}$$

où Ω est la fréquence de Bark calculée à partir de la fréquence angulaire ω par la définition :

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi}\right)^2 + 1\right)^{\frac{1}{2}}\right) \tag{3.5}$$

Le spectre de puissance du signal $P(\omega)$ (pair et périodique) est convolué avec la courbe de masquage:

$$\Theta(\Omega_k) = \sum_{\Omega=-1.3}^{\Omega=2.3} P(\Omega - \Omega_k)\Psi(\Omega) \qquad (3.6)$$

Puis, l'algorithme tente de faire l'approximation de la sensibilité de l'oreille humaine à différentes fréquences par l'intermédiaire d'une fonction de transfert $E(\omega)$:

$$\Xi(\Omega(\omega)) = E(\omega)\ \Theta(\Omega(\omega)) \qquad (3.7)$$

La non-linéarité entre l'intensité d'un son et son niveau de perception par l'oreille est réalisée en l'approchant par une loi de puissance :

$$\Phi(\Omega) = \Xi(\Omega)^{\frac{1}{3}} \qquad (3.8)$$

Enfin le spectre auditif est modélisé par un modèle tout-pôle. Une transformée de Fourier inverse discrète est appliquée sur le spectre auditif $\Phi(\Omega)$ pour obtenir les valeurs d'autocorrélation. $M+1$ premiers coefficients d'autocorrélation sont utilisés pour calculer les coefficients auto régressifs du modèle tout-pôle d'ordre M qu'on appelle les coefficients PLPs. Comme la méthode LPC, les coefficients cepstraux peuvent être obtenus à partir des coefficients PLPs.

3.3.1.3 Les coefficients MFCC (Mel-scale Frequency Cepstral Coefficients)

Les coefficients MFCC [Davis & Mermelstein, 1980] sont un autre type de coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole. Le codage MFCC utilise une échelle fréquentielle non-linéaire.

La fréquence mel-échelle est définie par :

$$B(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \qquad (3.9)$$

où f est la fréquence en Hz, $B(f)$ est la fréquence mel-échelle de f.

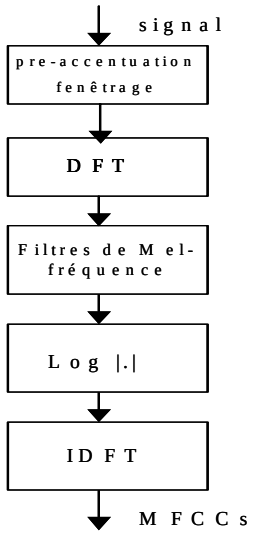


Figure 3.5 Calcul des MFCCs

Soit un signal discret {*s[n]*} avec $0 \le n \le N-1$, *N* est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète *S[k]* est obtenue :

$$S[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \qquad avec \quad 0 \le k < N \tag{3.10}$$

Le spectre du signal est multiplié avec des filtres triangulaires (figure **3.6**) dont les bande-passantes sont équivalentes en domaine mel-fréquence. Les points frontières *B[m]* des filtres en mel-fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1} \qquad 0 \le m \le M+1 \tag{3.11}$$

où *M* est le nombre de filtres, f_h est la fréquence la plus haute et f_l est la fréquence la plus basse pour le traitement du signal.

Dans le domaine fréquentiel, les points *f[m]* discrets correspondants sont calculés par l'équation :

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{3.12}$$

où B^{-1} est la transformée de mel-fréquence en fréquence. $B^{-1}(b) = 700 * (10^{b/2595} - 1)$

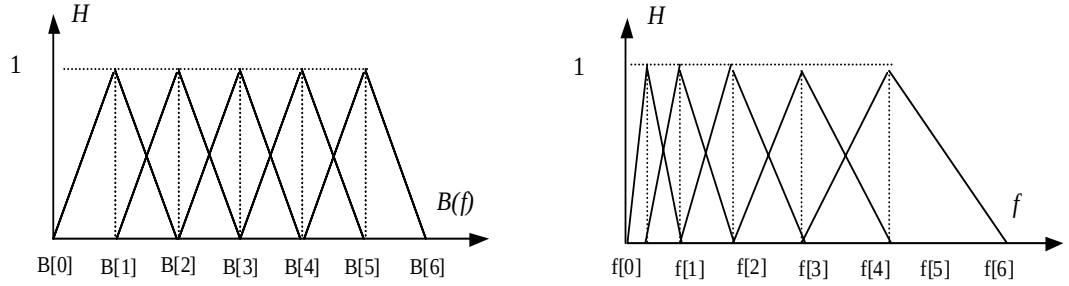


Figure 3.6 Les filtres triangulaires passe-bande en Mel-Fréquence (B(f)) et en fréquence (f)

Le coefficient $H_m[k]$ de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f[m-1] \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]} & si \quad f[m-1] \le k \le f[m] \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]} & si \quad f[m] \le k \le f[m+1] \\ 0 & si \quad k \ge f[m+1] \end{cases} \tag{3.13}$$

Pour un spectre lissé et stable, à la sortie des filtres un logarithme d'énergie (ou un logarithme de spectre d'amplitude) est calculé :

$$E[m] = \log\left[\sum_{k=0}^{N-1} |S[k]|^2 H_m[k]\right] \quad 0 \le m < M \tag{3.14}$$

Les coefficients cepstraux de mel-fréquence (MFCCs) peuvent être obtenus par une transformée de Fourier inverse à partir des coefficients aux sorties des filtres. Mais le nombre de MFCCs est moins grand que le nombre des filtres, donc une transformée de cosinus discrète est plutôt utilisée :

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n(m + \frac{1}{2})}{M}\right) \qquad 0 \le n < M \tag{3.15}$$

3.3.2 Reconnaissance de la parole

Les systèmes actuels de reconnaissance de la parole (RAP) sont fondés sur les principes de la reconnaissance de forme statistique [Jelinek, 1976],[Young,1996].

Soit les **m** mots prononcés dont les observations acoustiques sont $X = x_1, x_2, ..., x_T$. La tâche du système de RAP est d'estimer parmi toutes les séquences de mots M possibles la série $\hat{M}$ la plus probable connaissant X :

$$\hat{M} = Arg \max_M P(M / X) \tag{3.16}$$

Grâce à la règle de Bayes, il est possible d'écrire:

$$P(M / X) = \frac{P(X / M) P(M)}{P(X)} \tag{3.17}$$

Avec:

- *P(M/X)* est la probabilité *a posteriori* de la séquence M étant donnée une séquence des observations *X*.
- *P(X/M)* est la probabilité *a posteriori* d'émission des observations acoustiques *X* pour une séquence *M* donnée.
- *P(M)* est la probabilité *a priori* d'occurrence de la séquence *M*.
- *P(X)* est la probabilité d'occurrence des observations *X*. Elle est indépendante de M. *P(X)* est donc constant quand M varie.

P(X) ne dépend pas de *M* dans l'équation précédente, on peut écrire :

$$\hat{M} = \underset{M}{Arg\max}\, P(M / X) = \underset{M}{Arg\max}\, P(M)P(X|M) \qquad (3.18)$$

où P(M) est calculée par le modèle de langage. *P(X|M)* est calculée par le modèle acoustique.

3.3.2.1 Modèle de Markov caché

La technique essentiellement utilisée pour la modélisation acoustique dans les systèmes de reconnaissance de la parole est bien connue depuis vingt ans : c'est le modèle de Markov caché (HMM). Plus particulièrement, les HMMs utilisés en parole sont les HMMs d'ordre 1. Nous allons revoir les bases nécessaires à l'utilisation des ces modèles pour la reconnaissance de la parole dont les détails se trouvent dans [Rabiner, 1989].

Un HMM d'ordre 1 est caractérisé par un ensemble de paramètres λ:

$$\lambda = N,M,A,B, \pi \qquad (3.19)$$

avec:

- $S = \{s_1, s_2, ...,s_N\}$ - L'ensemble des N états du modèle. On note q_t l'état à l'instant t ;
- $A = \{a_{ij}\}$ - La matrice de transition, a_{ij} est la probabilité de transition de l'état s_i à l'état s_j.

$$a_{ij} = P(q_t=s_j|q_{t-1}=s_i)$$

$$\sum_{i=1}^{N} a_{ij} = 1 \quad 1 \le j \le N \qquad (3.20)$$

On rappelle que la probabilité de transition de l'état s_i à l'état s_j en général est :

$P(q_t=s_j|q_{t-1}=s_i,q_{t-2}=s_k,....)$.

Pour un HMM d'ordre 1 $P(q_t=s_j|q_{t-1}=s_i,q_{t-2}=s_k,....)= P(q_t=s_j|q_{t-1}=s_i)$, c'est-à-dire qu'elle ne dépend que l'état en cours q_t et de l'état précédant q_{t-1}.

- $B = \{b_j(o)\}$ - L'ensemble des probabilités d'émission de l'observation o dans l'état s_j. On note x_t l'observation à l'instant t ;
  - Si l'ensemble des observations est défini comme $x_t \in \{o_1, o_2, ..., o_M\}$ avec M nombre fini de symboles d'observation, alors on obtient un HMM discret.

$$b_j(o_k) = P(x_t = o_k | q_t = s_j) \quad 1 \le k \le M$$
$$\sum_{k=1}^{M} b_j(o_k) = 1 \quad 1 \le j \le N \tag{3.21}$$

  - Si l'ensemble des observations est continu, c'est-à-dire $x_t \in \{R^d\}$ avec d le nombre de dimensions du vecteur x_n, alors le HMM est continu et $b_j(o)$ est supposé être de la forme d'une fonction de densité de probabilité de variable o. Habituellement, une fonction de mélange de gaussiennes est utilisée pour représenter la fonction de densité de probabilité, c'est-à-dire :

$$b_j(o) = \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^{M} c_{jk} b_{jk}(o) \tag{3.22}$$

    avec M est le nombre de gaussiennes de l'état s_j. $N(o,\mu_{jk},\Sigma_{jk})$ où $b_{jk}(o)$ est la distribution gaussienne du $k^{ème}$ mélange de l'état s_j laquelle est définie par le vecteur moyen μjk et la matrice de covariance Σ_{jk} dans l'état s_j. c_{jk} est le coefficient de pondération du $k^{ème}$ mélange qui satisfait la contrainte:

$$\sum_{k=1}^{M} c_{jk} = 1 \tag{3.23}$$

  - La fonction de densité de probabilité $b_j(o)$:

$$\int_{-\infty}^{+\infty} b_j(o)do = 1 \quad 1 \le j \le N \tag{3.24}$$

- $\boldsymbol{\pi} = \{\pi_i\}$ - La distribution de la probabilité de l'état initial :

$$\pi_i = P(q_1 = s_i)$$
$$\sum_{i=1}^{N} \pi_i = 1 \tag{3.25}$$

Soit un HMM, il existe alors trois problématiques à résoudre :

- Estimation des probabilités :
  soit un modèle λ et la séquence d'observations $X = x_1, x_2,..x_T$, comment calculer $P(X|\lambda)$, la probabilité de la séquence des observations étant donné le modèle λ ?
- Estimation du meilleur chemin :
  soit une séquence d'observations $X = x_1, x_2,..x_T$ et le modèle λ, comment estimer la séquence des états $Q = q_1, q_1,...q_T$ pour maximiser $P(X, Q|\lambda)$?
- Entraînement (ou apprentissage) :
  soit un ensemble d'observations X et un modèle λ, comment ajuster les paramètres du modèle λ pour maximiser la probabilité $P(X|\lambda)$?

Le problème de l'estimation des probabilités est résolu par l'algorithme *Avant-Arrière (Forward-Backward)* [Baum, 1972]. Le problème de l'estimation du meilleur chemin est résolu par l'algorithme de *Viterbi* [Viterbi, 1967],[Forney, 1973]. Enfin, le problème de l'entraînement est le plus difficile car il a été montré qu'on ne peut pas trouver en général les paramètres du modèle pour obtenir le maximum global, seuls les paramètres pour le maximiser en local peuvent être calculés. Le problème de l'entraînement peut être résolu grâce l'algorithme *Baum-Welch* [Rabiner, 1989].

3.3.2.2 Modélisation de la parole avec un HMM

Un modèle HMM peut être utilisé pour modéliser le signal de parole. Ce signal est supposé être formé par une séquence des segments stationnaires ou pseudo-stationnaires. Le signal est alors transformé en une séquence de vecteurs acoustiques (représentants les observations). Chaque segment est ensuite modélisé par un état HMM (figure **3.7**).

Les vecteurs acoustiques associés au même segment stationnaire sont supposés être produits par le même état HMM. Chaque état HMM est caractérisé par une distribution de probabilité des vecteurs acoustiques. La transition d'un segment à un autre segment du signal est modélisée par la transition entre les états, laquelle est supposée être instantanée et caractérisée par la probabilité de transition de l'état.

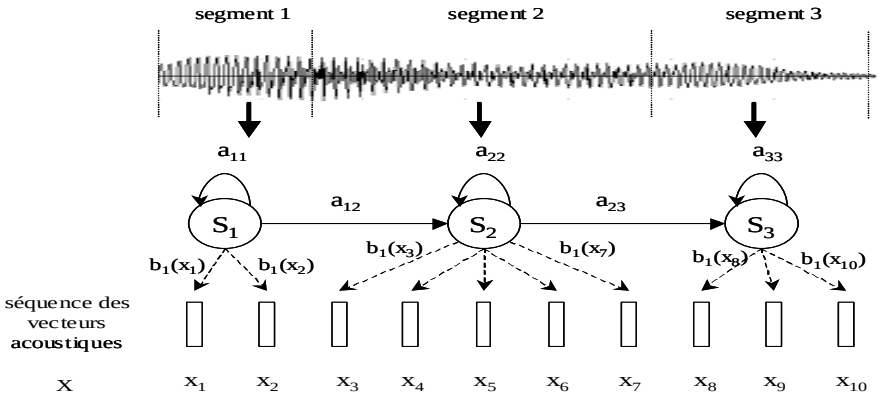


Figure 3.7 Un exemple de HMM 3 états modélisant un signal contenant 10 vecteurs acoustiques. Les trois segments stationnaires sont modélisés par trois états HMM. La séquence des états d'émission de la séquence $X=x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{10}$ contient $S=s_1\ s_1\ s_2\ s_2\ s_2\ s_2\ s_2\ s_3\ s_3\ s_3$.

3.3.2.3 Modèle de langage

Dans un système de reconnaissance de la parole, le modèle de langage est utilisé pour limiter le nombre des combinaisons entre les mots d'une phrase. Autrement dit, le modèle de langage calcule une probabilité pour toute séquence de mots. Un bon modèle de langage permet une augmentation du taux de reconnaissance et une réduction de la complexité de la procédure de recherche [Paescler & Ney, 1989]. Les modèles de langage qui sont utilisés essentiellement dans les systèmes de reconnaissance de la parole sont les modèles stochastiques *N-gramme*. Un modèle de langage N-grammes est fondé sur l'hypothèse que la probabilité d'un mot dans une phrase dépendant seulement de n-1 mots précédents.

Supposons que la séquence de mots possible $M = w_0, w_1, ...w_k, ..., w_K$ avec K variable qui représente le nombre de mots de la séquence M et w_0 qui désigne un état initial non associé à un mot du lexique, la probabilité $P(M)$ peut s'écrire :

$$\begin{aligned} P(M) &= P(w_0, w_1, ..., w_k, ..., w_K) \\ &= \prod_{k=1}^{K} P(w_k | w_{k-1}, w_{k-2}, ..., w_0) \end{aligned} \tag{3.26}$$

Pour le modèle N-grammes, l'équation devient:

$$\hat{P}(M) = \prod_{k=1}^{K} P(w_k | w_{k-1},..., w_{k-N}) \quad (3.27)$$

Les probabilités $P(w_k|w_{k-1},...,w_{k-N})$ peuvent être calculées à partir d'un grand corpus de texte.

$$P(w_k \Big| w_{k-1},..., w_{k-N}) = \frac{\mathrm{N}_{k-N}^{k}}{\mathrm{N}_k} \quad (3.28)$$

avec N^k_{k-N} représentant le nombre de fois que la séquence $(w_k, w_{k-1},...,w_{k-N})$ a été observée et N_k représentant le nombre de fois que le mot w_k est apparu dans la base d'entraînement du modèle de langage.

Un modèle bi-grammes ($N=2$) ou trigrammes ($N=3$) est utilisé habituellement dans les systèmes de reconnaissance de la parole [Young, 1996].

3.4 Reconnaissance de la langue tonale

Comme nous l'avons déjà écrit dans les paragraphes précédents, le modèle de production du signal de parole peut être représenté par un modèle simplifié qui présente deux parties : un filtre et une source d'excitation qui peut être soit une source d'impulsions quasi périodiques ou soit une source de bruit aléatoire. Le filtre modélise les caractéristiques du conduit vocal. La source d'excitation pseudo-périodique modélise la vibration des cordes vocales.

Dans un système de reconnaissance de la parole tel que nous venons de le décrire, les informations du vecteur acoustique utilisé pour caractériser les segments de parole ne concernent que le filtre. Ils ne tiennent pas compte de la source d'excitation. Cependant, ces systèmes donnent d'excellents résultats pour les langues occidentales.

Dans les langues tonales comme le Mandarin, le Thaï ou le Vietnamien, comme nous l'avons décrit dans notre chapitre 2 portant sur la phonologie et la phonétique de la langue vietnamienne, une syllabe peut être produite avec différents tons. Changer le ton, revient alors à changer la signification de la syllabe, c'est-à-dire pour ces langues mono-syllabiques, à changer de mot.

Le système de reconnaissance d'une langue asiatique doit donc compter sur cette caractéristique tonale fondamentale de la langue. Nous allons présenter ci-dessous l'état de

l'art des systèmes de reconnaissance du Mandarin, car ceux-ci ont été étudiés depuis plusieurs années dans la communauté scientifique.

3.4.1 Reconnaissance de syllabes isolées

Une caractéristique linguistique particulière du Mandarin tient dans le fait que c'est une langue tonale dont les mots contiennent une ou plusieurs mono-syllabes. Le Mandarin présente environ 10.000 mots fondés sur 1345 mono-syllabes avec tons. Ces mono-syllabes sont elles-mêmes la combinaison de 408 syllabes indépendamment du ton (appelées base-syllabes) et de 4 tons lexicaux et un ton neutre. Une syllabe peut être décomposée en deux parties INITIALE/FINALE comme nous l'avons présenté dans le chapitre 2 de notre mémoire.

Un système a été développé par [Lee et al, 1993] pour la reconnaissance du Mandarin pour le mode d'utilisation « syllabe isolée ». Dans ce mode, la frontière entre les syllabes est facilement déterminée par les silences. La figure **3.8** présente le schéma principal du système. Au niveau acoustique le système comprend deux parties séparées :

- reconnaissance des tons
- reconnaissance des syllabes sans ton (base-syllabes).

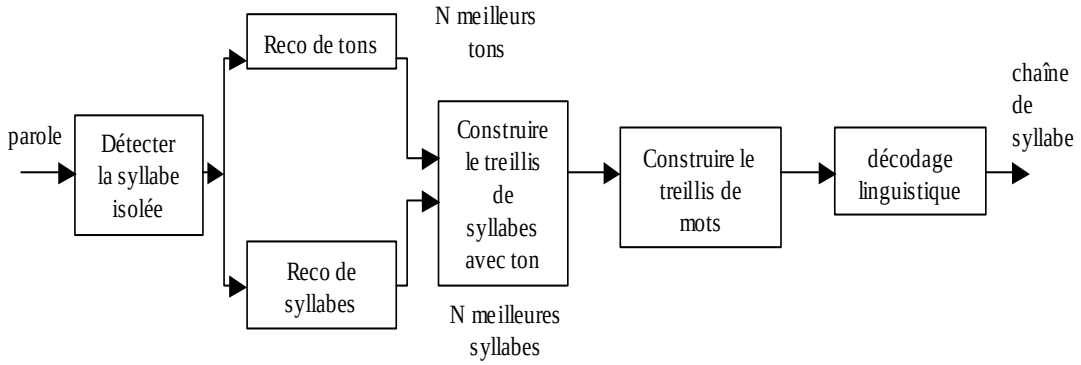


Figure 3.8 Schéma de principe de la reconnaissance du Mandarin (d'après [Lee et al, 1993])

3.4.1.1 Reconnaissance des syllabes sans ton :

Une syllabe comprend deux parties INITIALE/FINALE. La partie INITIALE est normalement plus courte que la partie FINALE. Les syllabes dont les parties FINALEs sont identiques et les parties INITIALEs différentes sont alors plus difficiles à distinguer.

La reconnaissance de syllabe sans tons est fondée sur deux hypothèses :

- la partie FINALE est indépendante de la partie INITIALE ;
- la partie INITIALE dépend de la partie FINALE.

Comme le mode d'utilisation du moteur de reconnaissance est celui de la reconnaissance de syllabes isolées, l'influence entre deux syllabes consécutives n'est pas prise en compte dans le système.

Avec ces hypothèses, les parties FINALEs sont modélisées par des modèles indépendants du contexte. Il y a 38 modèles de parties FINALEs. Les parties INITIALEs sont modélisées par des modèles dépendants du contexte à droite. Il y a 99 modèles de parties INITIALEs.

3.4.1.2 Reconnaissance de tons :

La reconnaissance de tons est fondée sur la reconnaissance de contour de pitch. Le Mandarin possède 5 tons. L'information portée par les tons est surtout présente dans la partie FINALE de la syllabe. En mode de syllabes isolées, on suppose que les tons sont indépendants du contexte. Les tons sont alors modélisés par 5 modèles indépendants du contexte.

3.4.1.3 Treillis de syllabes et treillis de mots

La syllabe avec ton est construite comme étant la combinaison d'une base-syllabe (syllabe indépendamment du ton) et d'un ton. Cependant, les deux processus de reconnaissance de tons et de reconnaissance de la base-syllabe, qui travaillent en parallèle dans le modèle de [Le et al. 1993], ne sont pas parfaits et ne produisent pas 100% de réussite. Pour améliorer le taux de reconnaissance du système complet, un modèle de langage est utilisé.

Pour une syllabe à reconnaître les deux processus de reconnaissance vont proposer chacun leurs N-meilleurs candidats, à partir desquels le système complet proposera les M-meilleurs

syllabes avec ton. Pour une chaîne de syllabes à l'entrée du système, un treillis de syllabes au niveau acoustique est construit. Puis un treillis de mots (un mot peut comprendre une, deux ou trois syllabes) est créé à partir d'un lexique.

3.4.1.4 Décodage linguistique

A partir du treillis de mots possibles, le processus de décodage linguistique va sortir une chaîne de mots meilleurs candidats fondés sur les contraintes syntaxique et stochastique. Dans les systèmes actuels à architecture N-grammes, les modèles bi-grammes ou trigrammes sont plus particulièrement utilisés. Une particularité du Mandarin est qu'une syllabe possède plusieurs caractères différents. De plus, un mot peut être constitué d'un caractère ou de plusieurs caractères. Chaque caractère a un sens linguistique en lui-même. C'est pour cette raison que le modèle *N-grammes de caractères* est utilisé plutôt que le modèle *N-grammes de mots* pour lequel les donnés d'entraînement sont limitées. Par contre, si les données d'apprentissage sont suffisantes, alors le modèle *N-grammes de mots* va donner de meilleurs résultats [Lee, 1997].

3.4.2 Reconnaissance de la parole continue

Les systèmes de reconnaissance de la parole continue en Mandarin en cours d'études, sont réalisés essentiellement sur la base de deux méthodes, parmi lesquelles l'une est le développement de la méthode de reconnaissance de syllabe isolée avec des adaptations pour le cas des mots continus, et l'autre est fondée sur les techniques bien maîtrisées de reconnaissance de langues occidentales comme l'anglais ou le français. Nous allons présenter les points principaux de ces deux méthodes.

3.4.2.1 Méthode "deux processus"

Avec cette méthode, de la même manière que pour la reconnaissance des syllabes isolées, une syllabe est construite à partir des résultats de deux processus séparés : la reconnaissance de la syllabe sans tons et la reconnaissance du ton.

Reconnaissance de tons :

Pour le cas des mots continus, la reconnaissance du ton devient difficile à cause des effets de la coarticulation, le ton dépendant alors du contexte. Le Mandarin contient seulement 5 tons (4 tons lexicaux et 1 ton neutre). Cependant, pour modéliser un ensemble de modèles dépendants de contexte, il est nécessaire de posséder 175 modèles [Wang et all, 1997], c'est-à-dire que 175 modèles = 5^3 (modèle dépendant du contexte gauche-droit pour le ton au milieu de la phrase) + 5^2 (modèle dépendant du contexte gauche pour le ton à la fin de la phrase) + 4*5 (modèle dépendant de contexte de droite pour le ton au début de la phrase, ton5 n'est jamais au début d'une phrase) + 5 (modèle isolé).

En fait en tenant compte des particularités acoustiques, le nombre des modèles peut être réduit aux 23 modèles au lieu de 175 modèles [Wang et al, 1995] [Wang et al, 1997]. Il a montré que dans le cas où les données d'entraînement sont limitées, la reconnaissance de tons avec 23 modèles dépendants de contexte donne un taux de reconnaissance 89.8 %, alors que celui en utilisant 175 modèles complets dépendants de contexte donne 88.2%. La technique utilisant 5 modèles indépendants de contexte donne quant à elle 86.9% [Wang et al, 1995]. Cependant si les données d'entraînement sont suffisantes, on peut supposer que l'utilisation des 175 modèles dépendants du contexte va donner un résultat meilleur que celui obtenu en utilisant uniquement 23 modèles.

Reconnaissance de base-syllabe :

Le Mandarin est une langue mono syllabique. Il n'y a pas de liaison entre les syllabes. Les effets de la coarticulation dans une syllabe ont plus de signification que ceux produits par la coarticulation entre deux syllabes. Cette particularité permet de réduire le nombre des modèles dépendants du contexte dans le cas où les données sont limitées.

Les unités acoustiques à reconnaître utilisées dans ce cas peuvent être INITIALEs et FINALEs. L'utilisation des parties INITIALEs dépendantes du contexte droit et des parties FINALEs indépendantes du contexte permet d'obtenir un taux de reconnaissance des base-syllabes de 88.2%.[Wang et al, 1995].

Une autre technique pour modéliser la syllabe consiste à utiliser des unités acoustiques plus petites, les PLUs (phone like units). Par l'utilisation des PLUs une partie INITIALE correspond alors à un seul PLU et une partie FINALE, plus longue, peut comprendre de une à

trois PLUs. Avec cette caractérisation, le Mandarin présente 33 PLUs. Pour modéliser les PLUs dépendantes du contexte, il faut constituer un ensemble complet de 511 modèles. Cependant, dans le but de limiter le besoin en données d'entraînement du système de reconnaissance, le nombre des modèles est diminué à 149 PLUs dépendants du contexte droit pour que chaque modèle soit entraîné par le plus de données possible [Lyu et al, 1995].

Reconnaissance de la syllabe avec ton :

Lorsque la reconnaissance de la base-syllabe et la reconnaissance du ton sont deux processus séparés, la reconnaissance de la syllabe, fondée sur la combinaison des deux résultats, devient problématique car les frontières entre les syllabes sont difficiles à déterminer exactement. Pour résoudre ce problème de synchronisation entre les deux processus de reconnaissance, un algorithme appelé CSM (Concatenated Syllable Matching) a été proposé par [Shen et al, 1994]. Cet algorithme a été par la suite développé pour synchroniser la reconnaissance des base-syllabes et la reconnaissance des tons [Wang et al, 1995] :

- la détermination des points de début de syllabe et des points de fin de syllabe dans une phrase est fondé sur l'analyse du contour d'énergie. Les creux du contour d'énergie marquent les débuts et les fins possibles des syllabes. La figure **3.9** donne un exemple des points de début et de fin d'une syllabe, ainsi que le contour d'énergie correspondant.

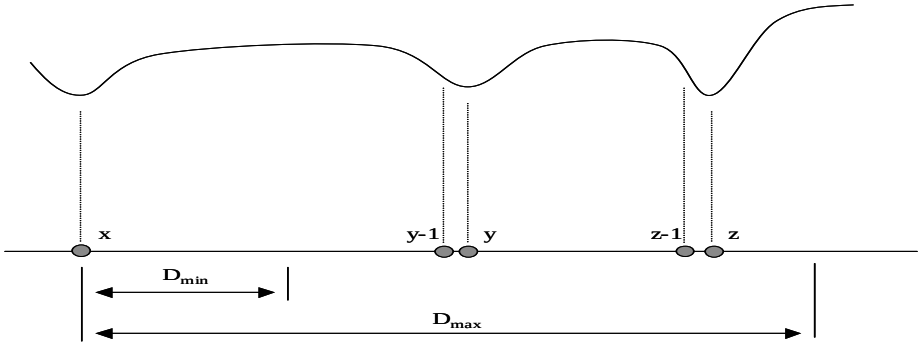


Figure 3.9 Un exemple du contour d'énergie et les points au début et à la fin de la syllabe.

Ici si *x* est le point au début de la syllabe, les points *y-1* et *z-1* sont les points possibles à la fin de la syllabe correspondant du point x. Les points *y* ou *z* représentent les points possibles pour le début de la syllabe suivante. En plus pour la syllabe dont le point terminal est le point *z-1*, les points *x* et *y* représentent les points possibles pour le début de la syllabe. Pour réduire les calculs, les points au début et à la fin des syllabes sont

cherchés dans un intervalle limité [D_{min}, D_{max}] avec D_{min} et D_{max} correspondant aux durées minimale et maximale d'une syllabe.

- Un score accumulé du point terminal y-1 de la syllabe est calculé par une méthode de programmation dynamique :

$$T[y-1] = \max_{x}\left\{T[x-1] + \max_{i}\left[S_{bs_i}(x, y-1) + S_{t_i}(x, y-1)\right]\right\} \tag{3.29}$$

où *T[y-1]* est le score accumulé du point terminal *y-1*de la syllabe, *T[x-1]* est le score accumulé du point terminal de la syllabe précédente, $S_{bsi}(x,y-1)$ et $S_{ti}(x,y-1)$ sont le score de la base-syllabe et le score du ton pour une syllabe *i* dans une section *(x,y-1)*.

- A la fin de la phrase, une chaîne de mots possibles peut être obtenue par une procédure de recherche arrière à partir du score accumulé le meilleur.

Comme dans le cas de reconnaissance de syllabes isolées, un treillis de syllabes va être produit à la fin de la phrase au lieu d'une chaîne de syllabes. Pour chaque section *(u,v)* les N syllabes les meilleures seront donc mémorisées au lieu d'une seule syllabe la meilleure.

3.4.2.2 Méthode à "un processus"

Dans la méthode utilisant un seul processus, la syllabe est reconnue immédiatement, sans distinction du ton ou de la base-syllabe. Afin de reconnaître la syllabe complète, des informations de pitch sont rajoutées au vecteur acoustique (contenant les coefficients cepstraux) habituellement utilisé dans la reconnaissance des langues occidentales.

Pour cela, un algorithme est proposé par [Chen et al, 1997][Chen et al, 2001]. La figure **3.10** présente le schéma de principe d'un système de reconnaissance utilisant la méthode à un « seul processus ».

Avec cette méthode, les unités acoustiques à reconnaître (syllabe, INITIALEs/FINALEs, phones) sont modélisées par des modèles dépendants du ton.

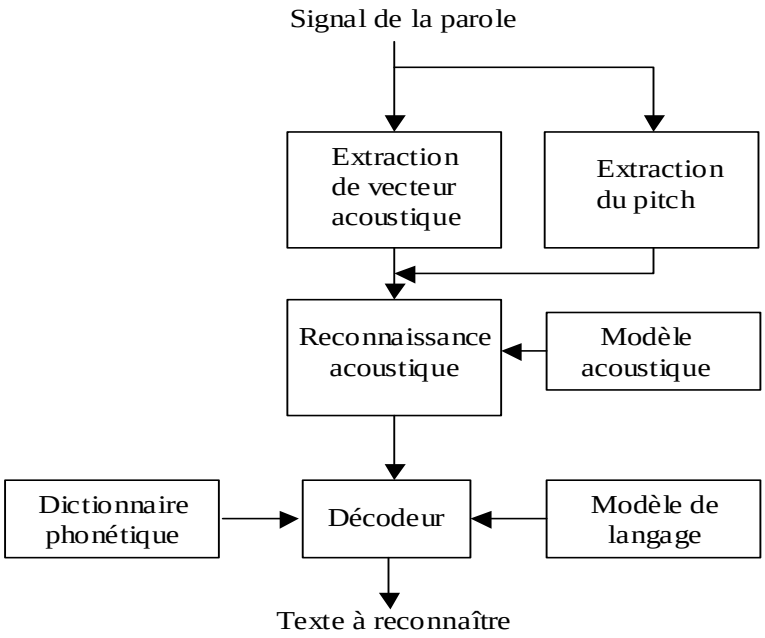


Figure 3.10 Système de reconnaissance de la parole pour les langues avec ton (d'après [*Chen* et al, 2001])

Méthode utilisant la structure INITIALEs/FINALEs :

L'ensemble des phonèmes utilisés comprend les modèles INITIALEs indépendants du ton et les modèles FINALEs dépendants du ton.

Méthode utilisant la structure "preme/toneme" :

Ici une hypothèse est proposée : le ton est centré essentiellement sur la partie (voyelle + son final). Avec cette hypothèse, un "preme" contient un son initial et un prétonal et est indépendant du ton, alors qu'un "toneme" est constitué d'une voyelle et d'un son final et dépend du ton.

Méthode utilisant l'information de pitch sur la voyelle (structure "voyelle principale"):

Cette méthode assume que l'information du ton n'est centrée essentiellement que sur la voyelle. Le but des méthodes utilisant la structure "preme/toneme" ou la structure "voyelle principale" est de diminuer le nombre des phonèmes utilisés. Les modèles sont appris sur plus de données. Ceci, dans le cas où les données sont limitées, améliore la robustesse du système.

La méthode utilisant la structure INITIALEs/FINALEs présente 215 phonèmes, la méthode utilisant la structure "preme/toneme" propose 158 phonèmes, alors que 73 phonèmes seulement sont utilisés dans la méthode utilisant la structure "voyelle principale".

3.4.3 Quelques commentaires sur les systèmes de reconnaissance du Mandarin

Pour la reconnaissance de la syllabe isolée, les systèmes proposés sont constitués de deux processus parallèles : reconnaissance du ton et reconnaissance de la base-syllabe. La reconnaissance de ton est fondée sur le contour du pitch.

Dans le cas de la reconnaissance de la parole continue, les résultats récents sont encore en cours d'analyse et de discussion.

Ces résultats du tableau 3.1 montrent que la valeur du taux d'erreur est réduite d'environ 4% (réduction relative du taux d'erreur ~ 24.9% par l'utilisation d'un vecteur acoustique comprenant l'information de pitch.

Cependant, dans le tableau 3.2, les conclusions semblent différentes : la valeur du taux d'erreur pour la méthode INITIALE/FINALE est seulement réduite de 0.4% par l'utilisation de l'information de pitch qui semble donc, dans ce cas, ne pas améliorer beaucoup le taux de reconnaissance. Autrement dit, il semble que les MFCCs contiennent des informations permettant de distinguer les syllabes entre les tons [Chang et al, 2000].

En fait, pour la parole continue l'application de la méthode "deux processus" est assez difficile puisqu'il existe des problèmes à résoudre : la frontière entre les syllabes et le contour de pitch dépendant beaucoup du contexte phonétique, la performance du système de reconnaissance des tons n'est pas bonne. La reconnaissance par la méthode "un processus" semble proposer plus d'avantages car les informations des tons sont traitées au même niveau que les paramètres cepstraux.

Méthode	Taux d'erreur (%)
preme / tonem vecteur acoustique ***avec*** l'information du pitch	13.65
voyelle principale vecteur acoustique comprenant l'information du pitch	12.61
modèle dépendant du ton vecteur acoustique ***sans*** l'information du pitch	16.79

Tableau 3.1 Résultats de reconnaissance du Mandarin (d'après [Chen et al, 2001])

Méthode	Taux d'erreur (%)
Initiale / Finale dépendant du ton MFCC vecteur acoustique ***sans*** l'information du pitch	6.43
Initial / Finale dépendant du ton MFCC vecteur acoustique ***avec*** l'information du pitch	6.03

Tableau 3.2 Taux d'erreur de reconnaissance avec la méthode Initiale/Finale (test proposé par Chang et al, 2000)

3.5 Conclusions

La reconnaissance de la parole est de plus en plus utilisée dans les applications demandant une interface homme-machine conviviale et naturelle. La technique utilisant des modèles HMM avec des vecteurs acoustiques MFCC, PLP ou LPC est le plus souvent utilisée dans les systèmes de reconnaissance actuels. Les systèmes de reconnaissance des langues occidentales (le français, l'anglais, etc.) sont fondés sur une caractérisation acoustique du signal de parole utilisant des paramètres cepstraux contenant uniquement des informations de type formantique. Par contre les systèmes de reconnaissance des langues tonales comme le Mandarin ont besoin d'informations supplémentaires caractérisant l'intonation, c'est-à-dire le pitch, afin de prendre en compte l'aspect lexical des variations du ton.

Pour la reconnaissance des langues tonales, en mode de mots isolés, la méthode utilisant "deux processus de reconnaissance" s'avère la plus performante puisque le nombre de modèles utilisés est le moins important. En mode de la parole continue, afin d'éviter les problèmes de caractérisation des frontières des mots, une méthode à un seul moteur global est utilisée, dans lequel l'information de pitch est traitée au même niveau que les paramètres

cepstraux. Cependant, le problème d'une utilisation efficace des informations de pitch n'est pas encore complètement résolu et les discussions sont encore ouvertes.

Notre but est de réaliser un système de reconnaissance du vietnamien en mode de mots isolés. Nous allons donc utiliser la méthode "deux processus". Ceci sera présenté dans les chapitres suivants.

3.6 Références

Atal B et Hanauer S (1971)
Speech analysis and synthesis by linear prediction of the speech wave
Journal of the Acoustic Soc. Am, vol 50, pp 637-655.

Baum L. E. (1972).
An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes
Inequalities, vol. 3, pp 1-8,

Boite R., Bourlard H., Dutoit T., Hancq J., Leich H. (2000)
Traitement de la parole
Presse Polytechniques et Universitaires Romandes.

Boitet C, Caelen J, Fafiotte, G, Keller, E, Lafourcade, M, Wehrli, E (1998)
"Integrating French within C-STAR II"
Grenoble,Report and demos of the CLIPS++ group

Calliope (1989)
La Parole et son Traitement Automatique
Editions Masson - Paris.

Chang E, Zhou J, Di S, Huang C, Lee K-F (2000)
Large vocabulary Mandarin speech recognition with different approaches in modeling tones
6th International Conference of Spoken Language Processing, Beijing

Chen C. J, Li H, Shen L, Fu G (2001)
Recognize tone languages using pitch information on the main vowel of each syllable
In Proc of ICASSP.

Chen C.J, Gopinath R. A, Monkowski M. D, Picheny M. A et Shen K (1997)
New methods in continuous Mandarin Speech Recognition
Eurpspeech'97, pp 1543-1546

Davis S.B., Mermelstein P. (1980)
Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,
IEEE Trans. on ASSP, vol. 28, no. 4, pp 357-366,.

Forney G. D. (1973)
Viterbi algorithm
Proccedings of the IEEE, vol. 61, pp 268-278.

Gales M.J.F (2001)
Adaptive Training for Robust ASR
Automatic Speech Recognition and Understanding Workshop, ASRU2001

Hermansky H. (1990)
Perceptual Linear Predictive (PLP) analysis of speech
Journal of the Acoustic Soc. Am, vol. 87, no. 4, pp 1738-1752.

Igounet S. (1998)
Eléments pour un système de reconnaissance automatique de la parole continue du français
Thèse Informatique, Université d'Avignon et des Pays de Vaucluse.

Itakura F et Saito S (1968)
Analysis synthesis telephony based upon the maximum likelihood method
In Kohasi Y, editor, 6th International Congress on Acoustics, Tokyo, pages C-5-5.

Jelinek F (1976)
Continuous Speech Recognition by Statistical Methods
Proc of the IEEE, vol 64, No 4, pp 532-557

Klatt D.H. (1977)
Review of the ARPA Speech Understanding Project
JASA, Vol. 62, n°6, pp.1345-1366

Lamel L. F, Adda G, Adda-Decker M (1996)
Les lexiques de prononciation dans les systèmes de reconnaissance de la parole
Séminaire GDR-PRC, Lexique et communication parlée, Toulouse, pp 1-10

Lee L-S, Tseng C-Y, Gu H-Y, Liu F-H, Chang C-H, Lin Y-H, Lee Y, Tu S-L, Hsieh S-H et Chen C- H (1993)
Golden Mandarin (I) - A Real Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary
IEEE Trans. on Speech and Audio Processing, vol. 1, no. 2, pp 158-179.

Lee L-S. (1997)
Voice Dictation of Mandarin Chinese
IEEE Signal Processing Magazine, pp 63-101.

Lyu R-Y, Chien L-F Hwang S-H, H H-Y, Yang R-C, Bai B-R, Weng J-C, Yang Y-J, Lin S-W, Chen K-J, Tseng C-Y, Lee L-S (1995)
Golden Mandarin (III) - A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary
In Proc of ICASSP, pp 57-60

Makhoul J. (1975)
Linear Prediction: A Tutorial Review
Proceedings of the IEEE, vol. 63, no. 4, pp 561-580.

Newell A., Barnett J., Forgie J.W., Green C.C., Klatt D.H.,. Licklider J.C.R, Munson J., Reddy D.R. & Woods W.A. (1973)
Speech Understanding Systems : Final Report of a Study Group
North-Holland/American Elseveir, Amsterdam.

Nguyen Thi H. L (1993)

Séparation aveugle de Sources à large bande dans un mélange convolutif: Application au rehaussement de la parole
Thèse doctorat de l'INP Grenoble

Paescler A et Ney H (1989)
Continuous speech recognition using a stochastic language model
In Proc of ICASSP, pp 699-702.

Picone J. W (1993)
Signal Modeling Techniques in Speech Recognition
Proc of the IEEE, vol. 81, No. 9, pp 1215-1247

Rabiner L. R, Juang B H (1993)
Fundamentals of Speech Recognition
published by Prentice Hall.

Rabiner L.R. (1989)
A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
Proceeding of the IEEE, vol. 77, no. 2, pp 257-284.

Shen J-L (1998)
Continuous Mandarin Speech Recognition for Chinese language with large vocabulary based on segmental probability model
IEE Proc-Vis. Image Signal Process, vol. 145, No. 5, pp 309-315.

Shen J-L, Wang H-M, Bai B-R et Lee L-S (1994)
An Initial Study on A Segmental Probability Model Approach to Large-Vocabulary Continuous Mandarin Speech Recognition
In Proc of ICASSP, pp 133-136

Spalanzani A (1999)
Algorithmes évolutionnaires pour l`étude de la robustesse des systèmes de reconnaissance automatique de la parole
Thèse, Université Joseph Fourier

Viterbi A.J. (1967)
Error bounds for convolutional codes and an asymptotically optimal decoding algorithm
IEEE Trans. Informat. Theory, vol. IT-13, pp 260-269.

Wang H-M, Ho T-H, Yang R-C, Shen J-L, Bai B-R, Hong J-C, Chen W-P, Yu T-L, Lee L-S (1997)
Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data
IEEE Trans on Speech and Audio Processing, Vol. 5, No. 2, pp195-200.

Wang H-M, Shen J-L, Yang Y-J, Tseng C-Y et Lee L-S (1995)
Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data
In Proc of LCASSP, pp 61-64

Young Steven., (1996)
A Review of Large-vocabulary Continuous-speech Recognition
IEEE Signal Processing Magazine, pp 45-57.

4

Corpus et caractérisation de tons

4.1 Introduction

Nous rappelons ce que nous avons présenté au chapitre 2 de notre mémoire. Le vietnamien est une langue monosyllabique (ou bisyllabique) avec des tons. Le vietnamien possède 6 tons. Au niveau de l'écriture, les tons ton2 à ton6 ont des symboles orthographiques correspondants, par contre le ton1 (ton plat) n'a pas de symbole orthographique.

Comme nous l'avons expliqué précédemment, les études sur la phonétique vietnamienne ont été réalisées dans les années 40 essentiellement. De plus, les études fondées sur des expérimentations pratiques sont peu nombreuses ces dernières années. Le principal objectif de notre travail de thèse étant de réaliser un moteur de reconnaissance du vietnamien, nous sommes donc contraints de combler en partie ce manque de données et nous présentons dans ce chapitre 4 notre travail de réalisation et d'analyse d'un corpus de mots isolés qui a été fait dans les buts suivants :

- déterminer les caractéristiques principales des tons vietnamiens en mode de syllabe isolée et sans contexte ;
- analyser et modéliser ces caractéristiques afin de les utiliser pour notre système de reconnaissance des tons pour un système de reconnaissance complet de la parole en langue vietnamienne.

4.2 Corpus

Pour réaliser cette première base de données sur la parole vietnamienne, nous avons défini un corpus afin qu'il puisse contenir l'essentiel des informations dont nous pensons avoir besoin pour la mise au point de notre système de reconnaissance. Notre corpus est donc construit avec les principaux objectifs suivants :

- réaliser la caractérisation acoustique et temporelle des 6 tons vietnamiens ;
- mettre au point un moteur de reconnaissance des tons ;
- concevoir ensuite un système de reconnaissance automatique complet de syllabe isolée pour des applications de commandes orales de processus. Pour notre application démonstrative, nous avons décidé d'utiliser notre système dans un contexte d'interface homme-machine utilisant les commandes orales pour manipuler les fonctions d'applications Windows ou d'applications Internet.
- de plus, afin que le corpus serve à des études ultérieures, en particulier sur la production des sons du vietnamiens, nous avons étendu le choix des mots prononcés afin d'obtenir un échantillon suffisant de toutes les voyelles vietnamiennes pour permettre leur analyse formantique complète.

C'est pourquoi, poursuivant les buts énoncés ci-dessus, nous avons décidé que notre corpus comprendrait des mots clés de commandes orales comme, par exemple, "ouvrir", "fermer", "imprimer", etc. mais aussi, pour la caractérisation des tons et leur reconnaissance, un ensemble de syllabes présentant les 16 voyelles et les 21 consonnes initiales. Chaque voyelle est combinée au minimum une fois avec chacun des 6 tons. Le nombre de syllabes du corpus comprend en définitive 131 syllabes.

Une particularité importante du vietnamien est que les différentes régions du pays n'utilisent pas le même accent. Il est courant de diviser le vietnamien en 3 dialectes principaux : dialecte du Nord, dialecte du Centre et dialecte du Sud. Notre corpus est composé de la manière suivante :

- 131 mots monosyllabiques correspondants aux 131 syllabes ;
- chaque mot est répété quatre fois par chaque locuteur. Les mots sont prononcés naturellement comme pour un dialogue en vie quotidienne (ce qui est différent de la

prononciation orthographique habituellement utilisée dans les écoles primaires pour l'apprentissage de la langue) ;

- le corpus est enregistré par 15 locuteurs, répartis en fonction du sexe mais aussi en fonction de l'origine géographique : 7 locuteurs (5 femmes et 2 hommes) du Nord, 4 locuteurs (2 femmes et 2 hommes) du Centre-Sud et 4 locuteurs (2 femmes et 2 hommes) du Sud.
- le corpus total correspond à 15*131*4 = 7860 items pour une durée d'environ 105 minutes de parole.

Les enregistrements ont été réalisés dans un studio isolé (ce n'est pas une chambre anéchoïque mais la qualité de l'environnement sonore est très correcte), à l'aide d'un microphone-casque Sennheiser HMD410. Ce microphone est d'excellente qualité avec une réponse en fréquence plate et optimisée pour la parole. De plus, l'utilisation d'un micro-casque permet de régler la position du microphone par rapport à la bouche, d'une manière quasi identique pour tous les locuteurs. L'acquisition du signal a été réalisée avec une précision de16 bits et une fréquence d'échantillonnage de 16 kHz.

Le logiciel d'enregistrement et de gestion de corpus que nous avons utilisé est le logiciel Emacop, spécialement conçu par notre équipe GEOD [Vaufrezdaz, 1998], que nous avons légèrement modifié pour qu'il accepte et affiche à l'écran les mots directement en langue vietnamienne avec une police de caractères vietnamiens standard.

mots		mots	
vietnamien	**français**	**vietnamien**	**français**
anh	grand frère	nhâm	neuvième signe du cycle décimal (de la cosmogonie ancienne)
ba	trois	nhân	multiplier
chia	diviser	nhan (đề)	titre
đưa	mener	nhăm	cinq (employé après les chiffres de dizaine)
ghi	enregistrer	phê	noter
hai	deux	sang	passer
in	imprimer	sau	derrière
khô	sec	to	grand
không	zéro	thay	remplacer
lăm	cinq (employé après les chiffres de dizaine)	thu	obtenir
lên	monter	thư	lettre
mơ	rêver	trang	page
mua	acheter	trên	supérieur
mươi	dizaine	xe	voiture
năm	cinq	xong	finir

Tableau 4.1 Les mots du corpus avec ton1

mots		mots	
vietnamien	**français**	**vietnamien**	**français**
bằng	égal	này	ce
bù	compenser	ngờ	douter
chè	thé	nhầm	tromper
đề	sujet	phòng	chambre
đồng	cuivre	trừ	moins
già	âgé	vì	car
mò	chercher à tâtons	vồ	saisir
mười	dix		

Tableau 4.2 Les mots du corpus avec ton2

mots		mots	
vietnamien	**français**	**vietnamien**	**français**
cũ	ancien	ngữ	moment
chĩa	braquer	nghĩ	penser
chữa	corriger	rẽ	tourner
đã	déjà	trễ	retard
đỡ	soulager	vẫn	toujours
đũa	baguettes	vỗ	taper
gõ	retaper		

Tableau 4.3 Les mots du corpus avec ton3

mots		mots	
vietnamien	**français**	**vietnamien**	**français**
bẩn	salle	khả (ái)	gentil
bảy	sept	lửa	feu
bẻ	forcer	mở	ouvrir
của	de	nghỉ	se reposer
chủ	patron	phải	droit
để	pour	thử	essayer
đổ	verser	trở	retourner
huỷ	annuler	vỏ	enveloppe

Tableau 4.4 Les mots du corpus avec ton4

mots		mots	
vietnamien	**français**	**vietnamien**	**français**
ấm	chaude	ngó	regarder
bé	petit	nhú	poindre
bóc	enlever	phá	détruire
bốn	quatre	rách	déchirer
cất	stocker	sáu	six
cắt	couper	số	nombre
chế	fabriquer	tám	huit
chép	copier	tớ	toi
chín	neuf	tới	venir
dưới	inférieur	thoát	quitter
đánh	frapper	trái	gauche
đến	arriver	xoá	effacer
đóng	fermer	xuống	descendre
gắp	retirer	xứ	pays
kiếm	chercher	ý	idée
mốt			

Tableau 4.5 Les mots du corpus avec ton5

mots		mots	
vietnamien	**français**	**vietnamien**	**français**
bận	occupé	lạ	inconnu
bịa	inventer	lại	venir
bộ	ensemble	mặt	face
cạnh	côté	một	un
cặp	couple	mục	rubrique
cộng	plus	nhẹ	léger
chợ	marché	sự	fait
dụ	séduire	tệp	fichier
đậm	gras	tựa	s'appuyer
đọng	croupir	trị	traiter
học	apprendre	trọ	loger
kệ	tant pis	vạch	tracer
kiện	colis		

Tableau 4.6 Les mots du corpus avec ton6

Les tableaux **4.1** à **4.6** présentent les mots du corpus, classés en fonction du ton utilisé (un tableau par ton). Le tableau **4.7** présente, quant à lui, le profil des 15 sujets locuteurs, qui sont pour la plupart des étudiants poursuivant leurs études universitaires à Grenoble et dont l'âge est compris dans un intervalle de 20 - 30 ans.

nom et prénom	code	sexe	age	village d'origine	situation actuelle
Phạm Ngọc Yến	PNY	f	40	Nord (Hanoi)	Nord (Hanoi)
Vũ Tuyết Trinh	VTT	f	25	Nord (Hanoi)	Nord (Hanoi)
Dương Phương Quỳnh	DPQ	f	25	Nord (Hanoi)	Nord (Hanoi)
Đỗ Hồng Hạnh	DHH	f	22	Nord (Hanoi)	Nord (Hanoi)
Doãn Hoàng Lan	DHL	f	22	Nord (Hanoi)	Nord (Hanoi)
Nguyễn Thị Huệ	NTH	f	23	Centre (Hue)	Centre (Hue)
Võ Thanh Huyền	VTH	f	22	Centre (Da nang)	Centre (Da nang)
Bùi Khánh Hằng	BKH	f	22	Sud (Ho Chi Minh)	Sud (Ho Chi Minh)
Lê Phúc Loan	LPL	f	22	Sud (Ho Chi Minh)	Sud (Ho Chi Minh)
Bùi Xuân Hồi	BXH	m	25	Nord (Hanoi)	Nord (Hanoi)
Tạ Tuấn Anh	TTA	m	25	Nord (Hanoi)	Nord (Hanoi)
Lê Việt Sỹ	LVS	m	22	Centre (Danang)	Centre (Danang)
Trần Viết Huân	TVH	m	27	Centre (Hue)	Sud (Ho Chi Minh)
Hồ Bảo Quốc	HBQ	m	35	Centre (Hue)	Sud (Ho Chi Minh)
Trần Thượng Tiến	TTT	m	27	Sud (Can tho)	Sud (Can tho)

Tableau 4.7 Les profils des sujets

4.3 Détection de la fréquence fondamentale

La fréquence fondamentale F0 (ou pitch) 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 F0, de l'intensité et de la durée, peut faire ressortir bien des caractéristiques du locuteur, comme son sexe, 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 (emphase), ou bien même pour caractériser les différences lexicales entre les mots.

Dans le traitement de la langue, une bonne connaissance de ces informations prosodiques est fondamentale, tant pour la synthèse, par exemple, où leur maîtrise permet de générer une parole synthétique plus naturelle et plus intelligible, que pour la reconnaissance où ces informations peuvent être des indices d'identification des éléments du signal de parole.

Pour les langues tonales comme le vietnamien, la variation de l'intonation sur la durée de la syllabe est appelée le ton et participe à différentier les mots entre eux, puisqu'une même syllabe prononcée avec des tons différents prendra alors des significations différentes.

4.3.1 Méthode de calcul du pitch

La méthode que nous avons appliquée est une méthode utilisant l'analyse en ondelettes pour estimer la période du pitch [Kadambe & Boudreaux-Bartels, 1992], [Wendt & Petropulu, 1996]. Nous avons choisi cette méthode car elle a les avantages suivants:

- Aucune hypothèse de stationnarité et quasi-stationnarité dans la fenêtre d'analyse;
- Capable d'estimer le pitch dans un intervalle de valeurs large;
- Permet d'estimer le pitch période par période.

Principe de la méthode:

- Pendant les segments voisés du signal de la parole, l'excitation principale du conduit vocal est un signal pseudo-périodique provenant de la vibration des cordes vocales. Les cordes vocales s'ouvrent sous l'effet de la pression des poumons, supérieure à la pression dans le conduit vocal, puis sous l'effet des forces élastiques des muscles et sous l'effet du passage de l'air au niveau de la glotte après leur ouverture, elles se referment brutalement. On peut représenter d'une manière très simplifiée ce signal d'excitation comme une suite d'impulsions. Le temps entre deux impulsions est appelé la période du pitch, la fréquence est appelée fréquence fondamentale F0.
- La fermeture des cordes vocales s'opère d'une manière plus brutale que l'ouverture. Ce phénomène correspond à un changement abrupt dans la forme d'onde du signal de la

parole. Pour estimer la période du pitch, l'instant de la fermeture de la glotte, qu'on appelle un « événement », est estimé et l'intervalle de temps entre deux événements consécutifs est mesuré. Nous allons appliquer ce principe de mesure, brièvement expliqué ici, en utilisant la transformée en ondelettes.

Les ondelettes sont des fonctions élémentaires avec lesquelles on va décomposer le signal x(t). Ces fonctions vont permettre une analyse temps-fréquence, et peuvent se déduire des filtres à Q constant. On peut également les introduire à partir d'une seule fonction ψ(t) appelée ondelette analysante. On construit ensuit les ondelettes $\psi_{a,b}(t)$ à partir de l'ondelette analysante ψ(t) par translation *b* et dilatation *a* (ou contraction) :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a}) \quad with \quad a,b \in \Re, \quad a \neq 0 \tag{4.1}$$

avec la condition:

$$\int_{-\infty}^{+\infty} \psi^2(t)dt < \infty \tag{4.2}$$

La transformée en ondelette continue est définie par :

$$CWT_x(a,b) = \int x(t)\psi^*{}_{a,b}(t)dt \tag{4.3}$$

On notera que la CWT (Continuous Wavelet Transform) convertit une fonction à une variable en une fonction à deux variables. La représentation d'une fonction par sa CWT est redondante et la transformée inverse n'est donc pas toujours unique. De plus toutes les fonctions CWT(a,b) ne sont pas forcément la CWT de la fonction x(t).

Si l'ondelette ψ satisfait la condition *d'admissibilité* :

$$C_\psi = \int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\omega)|}{\omega} d\omega < \infty \tag{4.4}$$

alors, la CWT admet une fonction inverse :

$$x(t) = \frac{1}{C_\psi} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} CWT(a,b)\psi_{a,b}(t)\frac{dadb}{a^2} \tag{4.5}$$

La transformée est souvent représentée par une image 2D en couleur ou en niveaux de gris correspondant au module et à la phase de la CWT(a,b).

Si le coefficient de dilatation *a*, qui est appelé l'échelle, est égal 2^j alors la transformée ondelette est dite dyadique :

$$D_yWT_x(b,2^j) = \frac{1}{2^j}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{2^j}\right)dt \qquad (4.6)$$

où D_yWT (Dyadic Wavelet Transform) est la transformée en ondelette dyadique.

Mallat et Zhong [Mallat & Zhong, 1989] ont indiqué que si on choisit une fonction ondelette lissée (une fonction lissée est une fonction dont la transformée de Fourier possède une énergie qui est centrée dans la région des basses fréquences), alors les maxima locaux de la transformée en ondelette dyadique D_yWT indiquent les variations brusques du signal. De même les minima locaux indiqueront les variations lentes du signal.

Kadambe et Boudreaux-Bartels [Kadambe & Boudreaux-Bartels,1992] ont utilisé cette propriété afin d'estimer l'instant de la fermeture de la glotte qui correspond à un changement abrupt du signal de la parole. Ils ont fait l'hypothèse que le changement abrupt concernant la fermeture de la glotte correspond à des maxima pour plusieurs échelles consécutives. Et ils ont utilisé trois échelles 2^3, 2^4, 2^5. A chaque trame correspondant à une échelle du signal, une valeur maximum est trouvée que l'on appelle le maximum global. Les maxima dont la valeur est égale ou supérieure à 0.8 fois le maximum global s'appellent des maxima locaux.

Si les localisations des maxima locaux de deux échelles consécutives sont identiques, alors le segment est voisé et le pitch est estimé. Si le maximum global est inférieur à un seuil quelconque, alors la trame est détectée comme une trame non voisée.

Se basant sur l'idée de Kadambe et Bordeaux-Bartels utilisant l'ondelette pour détecter le pitch, Wendt et Petropulu ont proposé d'utiliser une seule fonction à filtrer, c'est-à-dire d'utiliser une seule échelle où la bande passante de fréquence comprends l'intervalle de variation du pitch entre 30Hz - 500Hz [Wendt & Petropulu, 1996]. Les maxima locaux sont estimés avec cette seule échelle : si le maximum global est inférieur à un seuil, alors la trame est non-voisée ; dans le cas contraire, le pitch est estimé à partir des intervalles entre les maxima locaux.

4.3.2 Réalisation

Nous avons choisi la méthode proposée par Wendt et Petropulu pour implanter une fonction de calcul du pitch qui servira d'algorithme de base du système de reconnaissance des tons vietnamiens. La figure **4.1** présente l'organigramme de cette procédure de calcul du pitch.

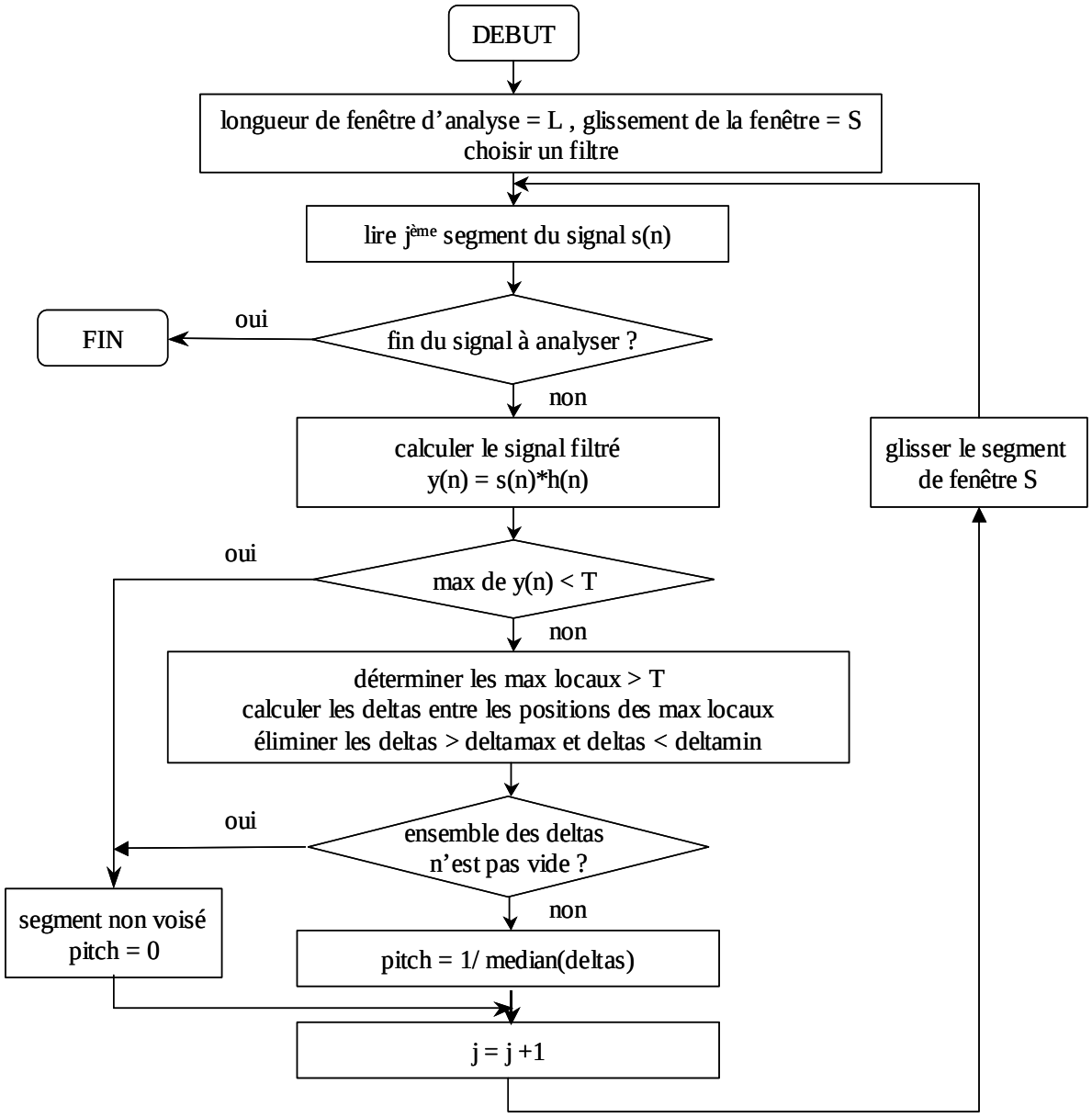


Figure 4.1 Schéma d'implantation de l'algorithme de [Wendt et al, 1996]

D'abord une fonction de filtre *h(n)* est choisie à partir de la fonction d'ondelette analysante $\psi_{ka}(n)$ (ou l'ondelette mère) et la fonction échelle $\varphi_{kb}(n)$ avec *ka* et *kb* les échelles

correspondant à des approximations aux fréquences limitées du pitch f_{max} et f_{min} ; $h(n)$ égale à la convolution de ces deux fonctions :

$$h(n) = \psi_{ka}(n) * \varphi_{kb}(n) \qquad (4.7)$$

Le signal à analyser est divisé en segments $s(n)$. Pour chaque segment, on calcule le signal filtré $y(n)$ de $s(n)$ avec le filtre $h(n)$. Si la valeur de $y(n)$ est inférieure à un seuil T, alors le segment est non voisé. Sinon les maxima locaux seront déterminés. Les deltas entre ses maxima sont aussi calculés. Une vérification pour les maxima est réalisée pour éliminer les deltas qui sont en dehors d'un intervalle compris entre deltamin et deltamax (avec deltamin = $1/f_{max}$ et delta max = $1/f_{min}$), c'est-à-dire les limites de périodes du pitch.

Si ce processus de calcul des deltas ne trouve pas de solution, alors le segment va être classifié comme non voisé. Sinon le pitch est calculé à partir des deltas.

Nous avons utilisé l'ondelette bi-orthogonale pour créer le filtre $h(n)$. La figure **4.2** présente respectivement:

- la fonction échelle $\varphi_{kb}(n)$ avec $kb = 4$;
- l'ondelette analysante $\psi_{ka}(n)$ à l'échelle $ka = 6$;
- le filtre $h(n)$.

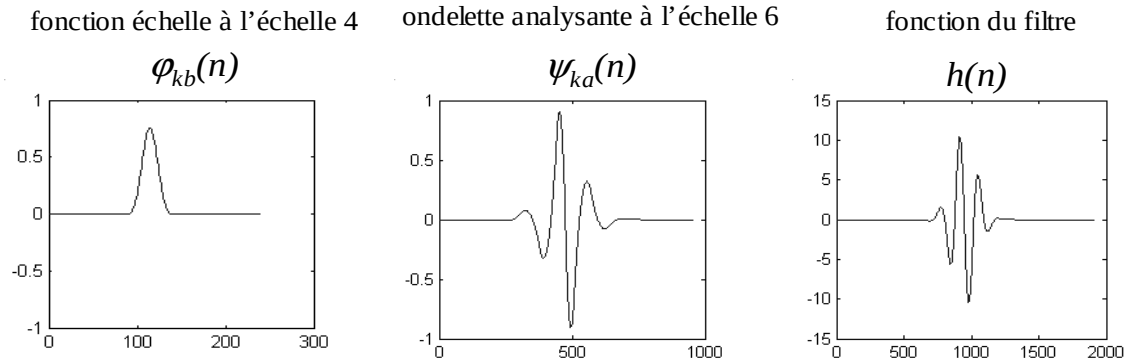


Figure 4.2 Un exemple de la forme de la fonction du filtre

4.3.3 Evaluation

4.3.3.1 Paramètres à comparer

Un algorithme de la détermination du pitch est évalué habituellement par deux aspects: la précision d'estimation du pitch et la décision de classification en segment voisé ou non-voisé [Rabiner et al. 1976].

Soient *Fx* le contour du pitch de référence et *F0* le contour du pitch estimé par l'algorithme :

- à l'instant *t* si les valeurs F_x et F_0 sont égales à zéro, alors le signal est classifié correctement comme le silence ou le non-voisé. Dans ce cas il n'y a pas d'erreur de classification voisée - non voisée.
- à l'instant *t* si la valeur de F_0 est non nulle mais celle de F_x est nulle, alors le signal non voisé (ou le silence) est classifié comme le signal voisé par l'algorithme. Dans ce cas il y a une erreur de classification des non-voisées.
- à l'instant *t* si F_x est non nulle mais F_0 est nulle, alors le signal voisé a été mal classifié incorrecte comme le signal non-voisée ou comme le silence. Dans ce cas il y a une erreur de classification des voisées
- à l'instant *t* si les deux valeurs *Fx* et F_0 sont non nulles, alors le signal voisée est classifiée correctement. La précision d'estimation du pitch de la méthode est alors évaluée :

Si $\frac{F_x - F_0}{F_x} \geq 0.2$	probablement la méthode estime la moitié de la valeur du pitch qu'on appelle l'erreur basse.
Si $\frac{F_x - F_0}{F_x} \leq -0.2$	probablement la méthode estime le double de la valeur du pitch qu'on appelle l'erreur haute
Si $\frac{\lvert F_x - F_0 \rvert}{F_x} \leq 0.2$	la valeur du pitch est acceptable

Soit un corpus du signal de référence correspondant aux contours du pitch de référence F_x . Le total de temps du corpus est *T*. Le total de temps des régions voisées du corpus est T_v et le total de temps des régions non-voisées ou des silences du corpus est T_{nv}. Alors $T = T_v + T_{nv}$. Les contours du pitch F_0 du corpus sont estimés par l'algorithme. Par la comparaison les deux contours F_x et F_0 on obtient :

- T_{nvf}: total de temps des régions non-voisées qui sont classifiées comme voisés;
- T_{vf}: total de temps des régions voisées qui sont classifiées comme non-voisées;
- T_{vgm}: total de temps des régions voisées avec les erreurs basses;
- T_{vgd}: total de temps des régions voisées avec les erreurs hautes;
- Valeur moyenne de la différence absolue et l'écart type entre F_x et F_0 pour les régions voisées qui sont classifiées comme voisés sans erreur par l'algorithme.

On définit:

- Erreur de classification des voisés : T_{vf} / T_v
- Erreur de classification des non-voisés : T_{nvf} / T_{nv}
- Erreur basse : $T_{vgm} / (T_v - T_{vf})$
- Erreur haute: $T_{vgd} / (T_v - T_{vf})$

4.3.3.2 Corpus de référence et évaluation

Nous avons utilisé le corpus de docteur Bagshaw [Bagshaw et al, 1993]. Le signal de parole est enregistré simultanément avec un microphone et un laryngographe dans une salle acoustiquement isolée. Le corpus comprend cinquante phrases. Chacune est prononcée par une femme et un homme. Les données de l'enregistrement du laryngographe sont utilisées pour calculer le contour de référence de pitch F_x.

Dans l'article de [Bagshaw et al, 1993] sont présentés les résultats d'évaluation de certaines méthodes de calcul du pitch. Ce sont:

- CPD (Cepstrum pitch determination) [Noll A.M, 1967].
- FBPT (Feature-based pitch tracker) [Phillips M.S, 1985].
- HPS (Harmonic product spectrum) [Schroeder M.R, 1968][Noll A.M, 1970].
- IPTA (Integrated pitch tracking algorithm) [Secrest B.G et Doddington, 1983].
- PP (Parallel processing method) [Gold. B et Rabiner. L, 1969].
- SRPD (Super resolution pitch determinator) [Medan Y, Yair. E et Chazan D, 1991].
- eSRPD (Enhanced version of SRPD) [Bagshaw P.C, Hiller S.M, Jack M.A, 1993]

Le tableau **4.8** présente les résultats de [Bagshaw et al, 1993],. Auxquels nous avons rajouté les résultats de la méthode de [Wendt et al, 1996] que nous avons implantée. Cette méthode

est appelée provisoirement wPDA (wavelet pitch determination algorithm). Les contours de pitch déterminés par les méthodes sont comparés avec le contour de référence Fx.

- *Classification de voisée et non voisée*: les résultats de la méthode wPDA sont comparables avec les autres algorithmes.
- *Erreur*: le total de l'erreur comprenant l'erreur haute et l'erreur basse est comparable avec la méthode eSRPD. La méthode ondelette obtient 1.26% d'erreur pour le corpus de l'homme et 0.8% pour le corpus de la femme par rapport à la méthode eRSPD qui obtient 1.46% d'erreur pour le corpus de l'homme et 0.63 % d'erreur pour le corpus de la femme.
- Pour la déviation absolue les résultats des méthodes sont similaires.

Les résultats de la méthode wPDA sont comparables avec les autres méthodes. Nous l'utiliserons donc pour notre travail.

	ADP	erreur de classification non-voisée (%)	erreur de classification voisée (%)	erreur (%)		déviation absolue (Hz)	
				haute	basse	moyen	écart type
corpus de la voix de l'homme	CPD	18.11	19.89	4.09	0.64	2.94	3.60
	FBPT	3.73	13.90	1.27	0.64	1.86	2.89
	HPS	14.11	7.07	5.34	28.15	3.25	3.21
	IPTA	9.78	17.45	1.40	0.83	2.67	3.37
	PP	7.69	15.82	0.22	1.74	2.64	3.01
	SRPD	4.05	15.78	0.62	2.01	1.78	2.46
	eSRPD	4.63	12.07	0.90	0.56	1.40	1.74
	wPDA	14.85	5.91	0.25	1.01	1.99	3.04
corpus de la voix de la femme	CPD	31.53	22.22	0.61	3.97	6.39	7.61
	FBPT	3.61	12.16	0.60	3.55	5.40	7.03
	HPS	19.10	21.06	0.46	1.61	4.59	5.31
	IPTA	5.70	15.93	0.53	3.12	4.38	5.35
	PP	6.15	13.01	0.26	3.20	6.11	6.45
	SRPD	2.35	12.16	0.39	5.56	4.14	5.51
	eSRPD	2.73	9.13	0.43	0.23	4.17	5.13
	wPDA	7.03	6.54	0.60	0.20	4.45	5.99

Tableau 4.8 Evaluation de l'algorithme de détermination du pitch.

4.4 Caractérisation des tons

Les caractéristiques principales du ton sont : le contour, le registre et la durée. Dans une syllabe, le ton est centré essentiellement sur la partie finale qui comprend le prétonal, la voyelle et le son final.

Pour chaque ton nous allons mesurer les valeurs du pitch des points caractéristiques du ton et calculer les statistiques comme la moyenne et l'écart-type. La durée du ton, c'est à dire la durée de la partie finale, est aussi déterminée. A partir des statistiques nous allons déduire les gabarits des tons pour les hommes et les femmes. Les gabarits seront construits à partir des valeurs maximales et minimales observées parmi les locuteurs pour chaque point de mesure de chaque ton.

4.4.1. Caractéristiques acoustiques des tons par les expérimentations pratiques

4.4.1.1 Ton1

Le ton1 est un ton monotone. Les points à mesurer sont le point initial (B) et le point final (E), il n'est en effet pas nécessaire de mesurer des points intermédiaires qui dans ce cas n'apportent pas d'information supplémentaire sur le contour. La figure **4.3** montre un exemple de mesure pour le sujet PNY, avec les points particuliers référencés.

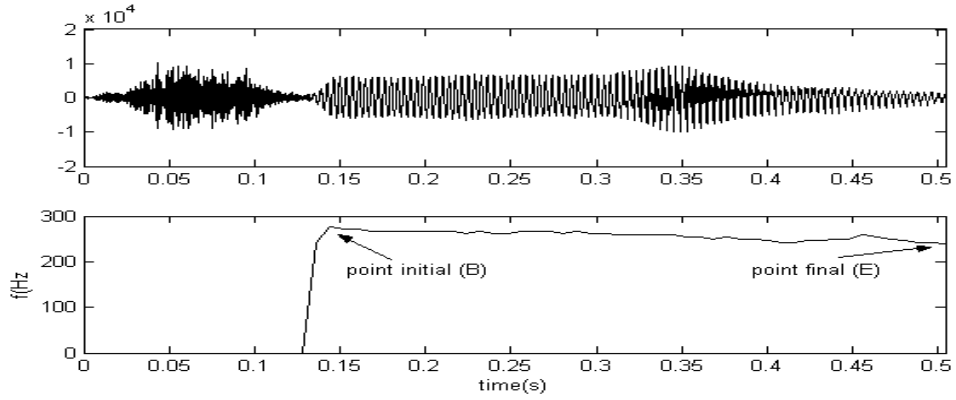


Figure 4.3 Un exemple du ton1 du sujet féminin PNY avec la syllabe "chia"

Le tableau **4.9** résume les valeurs du pitch obtenues sur tout le corpus pour le ton1 et la figures **4.4** donne les gabarits que nous avons déduits. Les gabarits du ton1 sont construits des valeurs maximales et minimales du point initial (B) et du point final (E) du ton1 de tous les locuteurs féminins ou masculins du corpus.

Nous pouvons constater que le ton1 est plat et descendant faible en général. La variation de fréquence Delta (la différence entre le point initial et le point final) est faible. L'intervalle de fréquence mesurée sur toutes les féminines est de 179 → 290Hz pour le point initial (B) et 183→301 Hz pour le point final (E). Pour les voix masculines, l'intervalle de B est 95→190 Hz et 93 → 188 Hz de E. La durée moyenne est 290 → 390 ms pour la voix féminine et 258→320 ms pour la voix masculine.

locuteur	fréquence (Hz)										durée	
	point initial				point final				delta			
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type
féminin			262→290	179 →225			228 →301	183 →225			290 →390	
PNY	255	12	285	225	246	12	271	213	-9	10	350	50
VTT	236	12	271	195	233	8	275	210	-3	12	340	40
DPQ	216	15	262	179	204	10	228	183	-12	13	330	40
DHH	250	14	280	207	240	9	262	225	-10	14	390	40
DHL	246	15	285	202	233	11	258	210	-13	11	290	30
NTH	241	14	280	210	256	13	301	225	15	14	334	45
VTH	250	14	290	216	251	12	275	216	0	14	370	66
BKH	242	11	280	216	231	11	261	205	-11	14	333	32
LPL	241	10	267	218	222	10	250	202	-18	9	312	40
masculin			120 →190	95 →130			117 →188	93 →135			258 →320	
BXH	128	7	149	110	127	5	142	113	-1	6	320	60
TTA	161	11	190	130	161	10	188	135	0	10	320	60
LVS	129	8	144	101	120	6	134	101	-9	8	319	49
TVH	104	4	120	95	101	13	117	93	-3	5	258	30
HBQ	145	7	164	130	145	5	158	125	0	7	291	45
TTT	130	5	145	117	123	5	141	109	-6	6	320	47

Tableau 4.9 Caractéristiques acoustiques du ton1

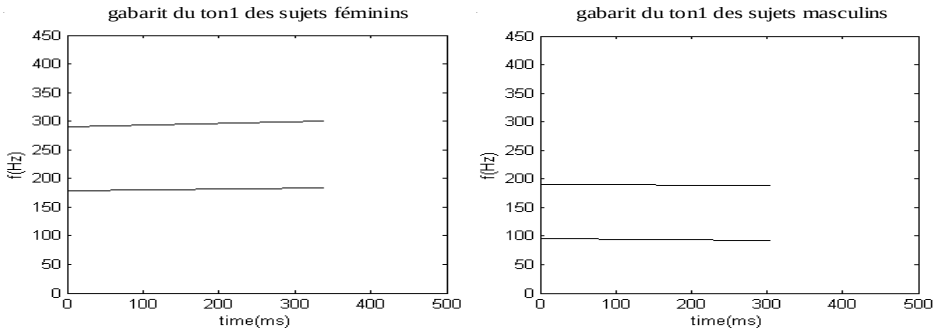


Figure 4.4 Gabarits du ton1

4.4.1.2 Ton2

Le ton2 est aussi un monotone. Le contour de ton2 est plat et descendant. Les points à mesurer est donc le point initial (B) et le point final (E). La figure **4.5** donne un exemple du ton2 avec les points à mesurer.

Le tableau **4.10** résume les valeurs du pitch des points à mesurer et la durée du ton2. La figure **4.6** présente les gabarits du ton2.

La fréquence du point initial est dans l'intervalle 166 → 266 Hz pour les voix féminines et 94 → 146Hz pour les voix masculines. Elle est plus basse que celle pour le ton1. Le delta entre le point final et le point initial est assez grand -17 → -44 Hz pour les voix féminines et -8 → -24 Hz pour les voix masculines. La durée est à peu près la même que celle du ton1.

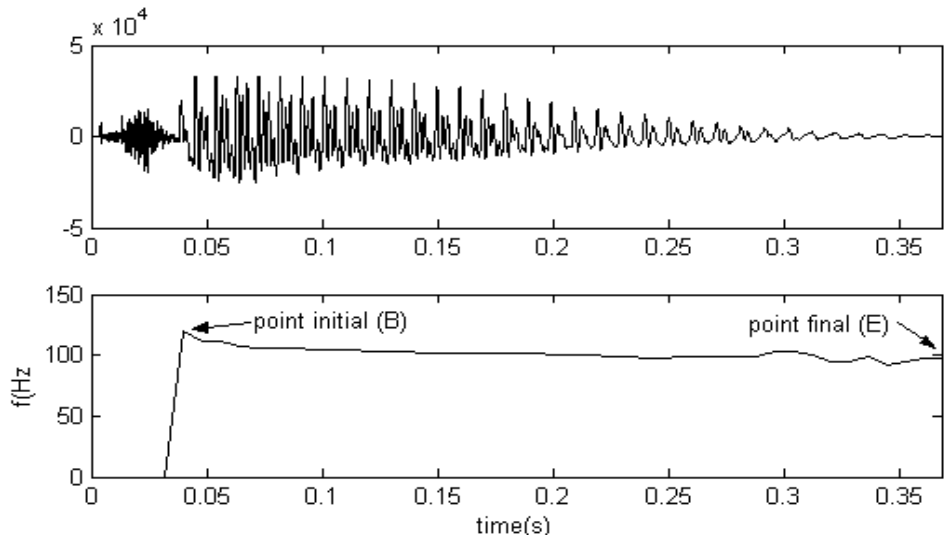


Figure 4.5 Un exemple du ton2 du sujet masculin LVS avec la syllabe "trừ"

locuteur	fréquence (Hz)										durée	
	point initial				point final				delta			
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type
féminin			228→266	166→210			172→219	129→179			295→405	
PNY	211	9	231	192	175	9	192	148	-36	10	360	50
VTT	203	11	231	177	186	9	205	164	-17	13	350	40
DPQ	185	12	228	166	154	7	172	129	-31	13	351	50
DHH	215	11	250	192	171	5	181	155	-44	11	405	40
DHL	214	13	252	190	176	7	188	155	-38	14	295	30
NTH	237	11	266	210	200	7	216	179	-36	11	375	42
VTH	209	8	228	197	179	9	219	163	-30	12	400	61
BKH	210	8	234	193	178	7	199	165	-31	10	367	30
LPL	211	9	240	194	173	9	196	155	-37	8	339	35
masculin			108→146	94→125			92→136	80→110			284→360	
BXH	110	5	121	98	101	5	114	91	-8	4	328	40
TTA	135	5	146	125	111	6	133	99	-24	6	314	40
LVS	110	6	128	97	96	4	107	86	-13	6	360	51
TVH	100	3	108	94	86	2	92	80	-13	3	284	28
HBQ	129	6	146	114	120	5	136	110	-8	5	310	41
TTT	112	4	126	103	93	4	108	85	-18	5	346	48

Tableau 4.10 Caractéristiques acoustiques du ton2

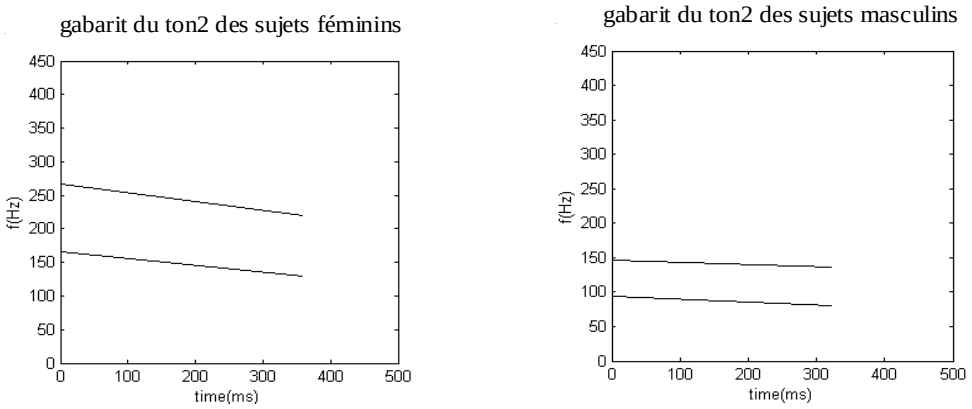


Figure 4.6 Gabarits du ton2

4.4.1.3 Ton3

Le mouvement de contour du ton3 est descendant et puis montant. Au milieu du ton, le contour présente une rupture. Une explication de ce phénomène est qu'il existe un mouvement de constriction glottale [Hoang T & Hoang M, 1975]. Mais celui-ci n'est pas obligatoire. C'est à dire que celui-ci dépend du locuteur. Le locuteur prononce le ton3 qui présente une rupture au milieu, d'autres fois il prononce le ton3 sans rupture. Ce phénomène existe essentiellement dans la façon de prononcer pour les sujets féminins plutôt que pour les sujets masculins. Le ton3 peut être divisé en 3 segments. Le premier segment et le troisième segment comprennent les segments monotones du contour. Le deuxième segment est le segment de la rupture. Le 1er segment est caractérisé par le point initial B1 et le point final E1 du segment. Le 3ème segment est caractérisé par le point initial B2 et le point final E2. Le 2ème segment ou le segment de la rupture est caractérisé par le point au milieu M entre les points E1 et B2.

La figure **4.7** donne un exemple du ton3 avec les points à mesurer.

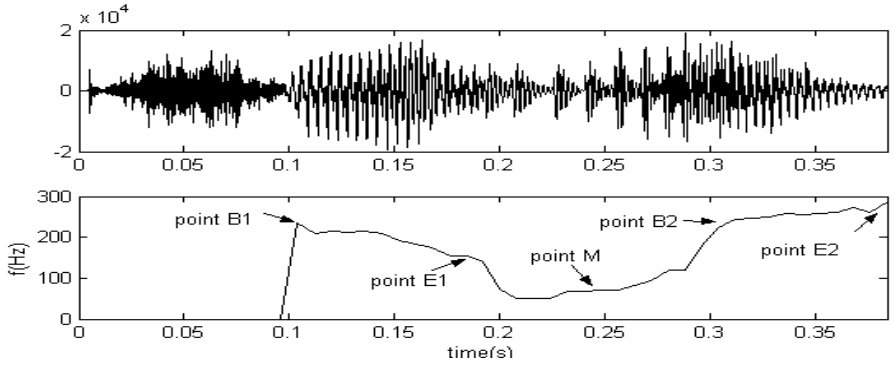


Figure 4.7 Un exemple du ton3 du sujet féminin DPQ avec la syllabe "chữa"

Le tableau **4.11** et la figure **4.8** présentent les caractéristiques et les gabarits du ton3.

Du point B1 au point E1, le contour est descendant. Du point B2 au point E2, le contour est montant. Le point initial B1 est plus bas que celui pour le ton1, Le point final E2 est plus haut que celui pour le ton1. Pour la voix féminine la fréquence du point final est 219 → 421 Hz et 103 → 302 Hz pour la voix masculine. Le delta entre les points final et initial est très grand 39 → 149 Hz pour la voix féminine et 16 → 93 Hz pour la voix masculine.

locuteurs	fréquence (Hz)																						durée (ms)	
	point B1				point E1				point M				point B2				point E2				delta		moyenne	écart type
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type		
féminin			232 285	170 213			189 267	101 182			165 253	0 55			224 333	84 203			296 421	219 333			237 358	
PNY	232	13	258	200	218	17	262	173	82	27	168	0	242	21	286	203	314	12	340	290	82	16	304	30
VTT	221	15	250	186	213	18	242	182	158	49	228	0	233	26	314	192	303	21	340	231	82	24	325	38
DPQ	201	15	242	177	194	24	232	118	109	49	213	0	210	39	258	111	275	30	340	219	71	31	275	31
DHH	224	16	266	190	205	17	258	170	103	37	216	0	251	28	308	182	373	20	421	333	149	23	357	27
DHL	243	18	285	213	227	18	267	179	132	48	253	0	255	33	333	199	344	29	410	285	100	30	237	41
NTH	224	10	246	200	196	17	235	158	119	49	225	55	204	45	276	107	264	17	296	229	39	18	348	33
VTH	209	10	232	184	165	40	213	101	97	32	205	52	180	60	271	84	305	18	340	266	96	16	334	42
BKH	200	17	245	170	159	13	189	135	85	40	178	0	190	13	224	163	292	22	356	261	91	25	358	38
LPL	201	11	236	176	186	8	203	164	58	34	165	0	209	16	272	180	277	25	375	246	75	21	325	31
masculin			108 175	88 129			99 174	66 131			86 162	0 42			102 241	56 157			131 302		106 213		257 363	
BXH	122	8	140	106	113	9	136	95	96	24	132	0	128	11	149	95	159	20	190	140	35	17	309	34
TTA	151	13	175	129	152	9	174	131	95	37	162	0	182	16	241	157	244	19	302	213	93	20	257	33
LVS	111	7	128	88	92	8	113	66	70	20	105	0	97	16	176	69	167	16	213	134	54	14	322	34
TVH	100	3	108	94	88	4	99	78	60	14	86	38	86	11	102	58	116	5	131	106	16	5	287	26
HBQ	119	6	133	110	104	5	122	90	64	29	153	0	110	29	161	56	161	12	190	128	42	12	312	36
TTT	101	4	113	90	89	6	101	74	76	15	96	42	95	13	172	79	168	13	200	134	66	13	363	44

Tableau 4.11 Les caractéristiques acoustiques du ton3

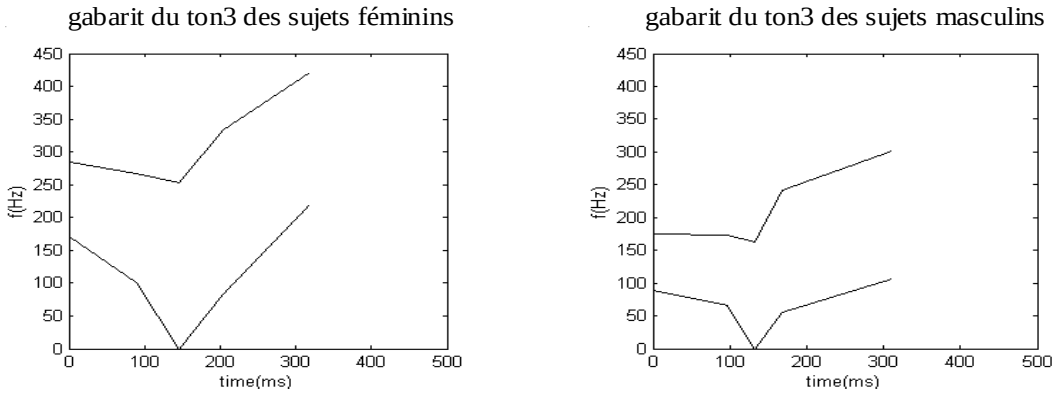


Figure 4.8 Gabarits du ton3

4.4.1.4 Ton4

Le ton4 a deux représentations différentes : le ton4 de la voix des locuteurs du Nord et le ton4 de la voix des locuteurs de la région centrale et du Sud.

Dans le cas des locuteurs du Nord, le contour du ton4 est descendant en général. Vers la fin, le contour peut être horizontal ou montant faible. Donc, les points à mesurer sont le point initial du ton B, le point M où la direction du contour est changée et le point final du ton E. La figure **4.9** donne un exemple du ton4 d'un sujet du Nord avec les points à mesurer.

Le tableau **4.12** et la figure **4.11** présentent les caractéristiques et les gabarits du ton4 des sujets du Nord. Le point initial est au même niveau que celui du ton2, la fréquence de point final est plus basse que pour le ton2. Le delta entre le point final et le point initial est plus grand que pour le ton2. Pour les voix féminines, le delta est -34 → -74 Hz et -18 → -35 Hz pour les voix masculines.

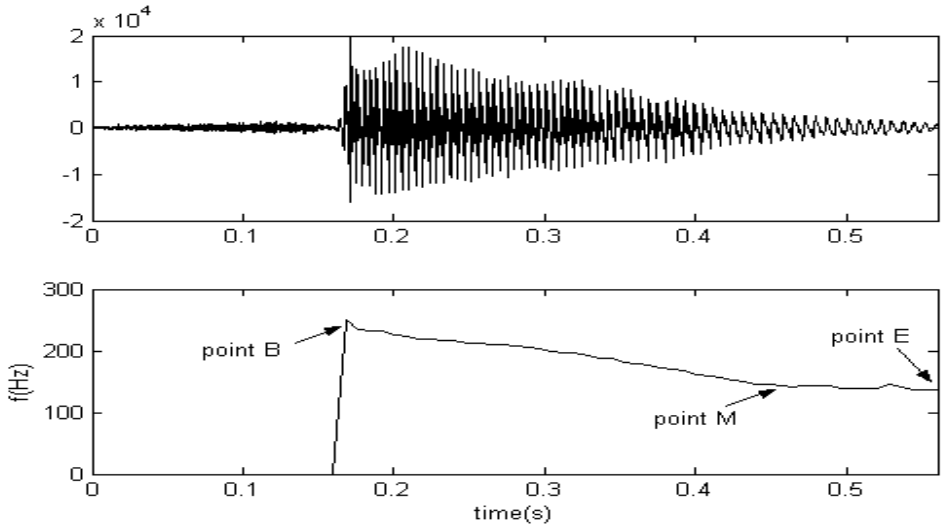


Figure 4.9 Un exemple du ton4 du sujet féminin DHH du Nord avec la syllabe "phải"

Dans le cas des locuteurs de la région centrale et du Sud, le contour du ton4 est semblable à celui du ton3. C'est à dire qu'on le prononce sans distinction du ton3. La figure **4.10** donne un exemple du ton4 d'un sujet du Sud avec les points à mesurer comme nous l'avons expliqué pour ceux du ton3. Le tableau **4.13** et la figure **4.12** présentent les caractéristiques et les gabarits du ton4 des sujets du Sud et du Central.

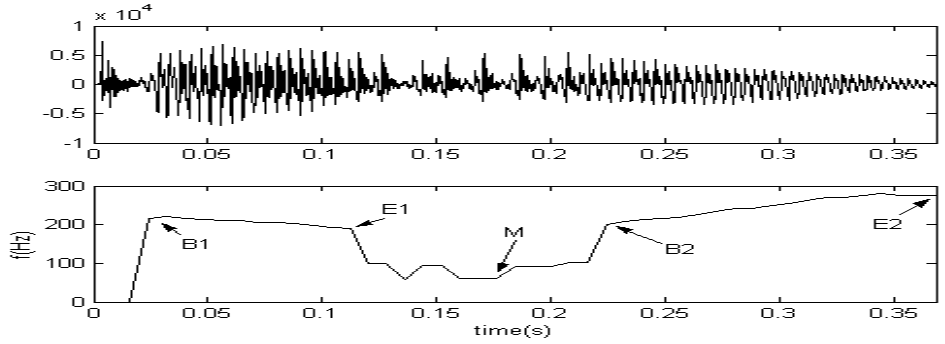


Figure 4.10 Un exemple du ton4 du sujet féminin LPL du Sud avec la syllabe "trở"

locuteur	**fréquence** (Hz)														**durée** (ms)			
	point B				**point M**				**point E**				**delta**		**B → M**		**B → E**	
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type	moyenne	écart type
féminin			210 → 271	150 → 188			148 → 195	101 → 136			163 → 216	101 → 136			184 → 243		206 → 314	
PNY	220	14	246	188	156	16	184	109	153	16	179	109	-67	19	226	50	257	60
VTT	207	13	231	179	169	12	195	136	173	15	216	136	-34	19	215	50	269	50
DPQ	180	16	210	150	129	8	148	101	129	9	163	101	-49	18	230	40	270	40
DHH	213	16	246	183	146	11	162	120	144	11	164	120	-69	21	243	50	314	40
DHL	220	18	271	186	148	13	178	105	146	13	177	105	-74	19	184	30	206	30
masculin																		
BXH	108	6	121	95	83	6	96	65	90	9	111	65	-18	11	205	50	280	50
TTA	135	7	156	118	102	4	116	93	100	5	118	73	-35	11	178	40	194	40

Tableau 4.12 Les caractéristiques acoustiques du ton4 des sujets du Nord.

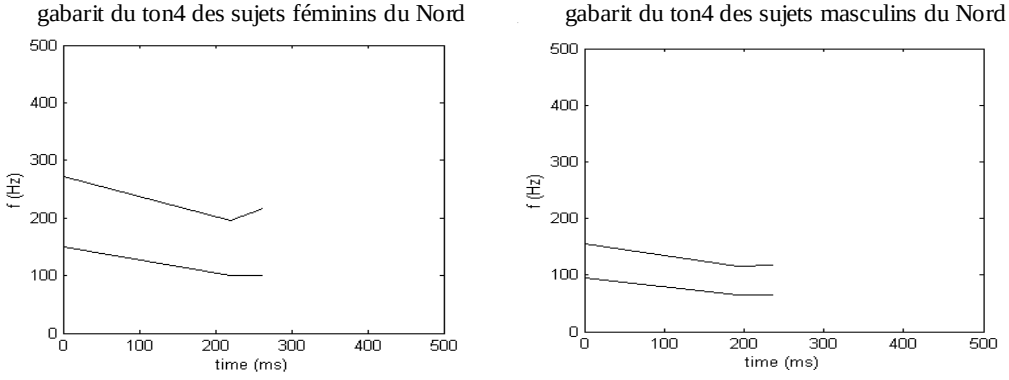


Figure 4.11 Gabarits du ton4 des sujets du Nord

locuteurs	fréquence (Hz)																						durée (ms)	
	point B1				point E1				point M				point B2				point E2				delta		moyenne	écart type
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type		
féminin																								
NTH	224	17	271	193	197	12	229	168	131	45	208	31	212	19	258	170	256	16	291	195	32	19	333	45
VTH	204	13	254	182	194	9	211	172	93	24	162	45	235	16	271	186	288	18	327	254	64	21	331	45
BKH	198	17	240	162	154	14	180	113	72	32	147	0	189	17	239	152	285	21	348	239	87	24	356	23
LPL	195	11	229	173	179	10	198	151	50	36	156	0	200	11	233	179	268	18	327	235	72	18	324	29
masculin																								
LVS	107	6	120	93	97	6	112	78	84	12	108	54	102	7	130	87	156	15	193	120	49	14	316	45
TVH	99	4	108	89	92	4	101	78	70	15	120	40	102	7	125	88	113	5	125	104	14	6	273	26
HBQ	119	8	145	104	105	4	117	96	36	33	107	0	142	14	179	106	166	11	197	138	46	16	288	31
TTT	102	6	122	87	91	6	105	73	81	15	102	40	97	8	117	70	162	14	188	129	60	12	342	42

Tableau 4.13 Caractéristiques acoustiques du ton4 des sujets du Sud et du Central

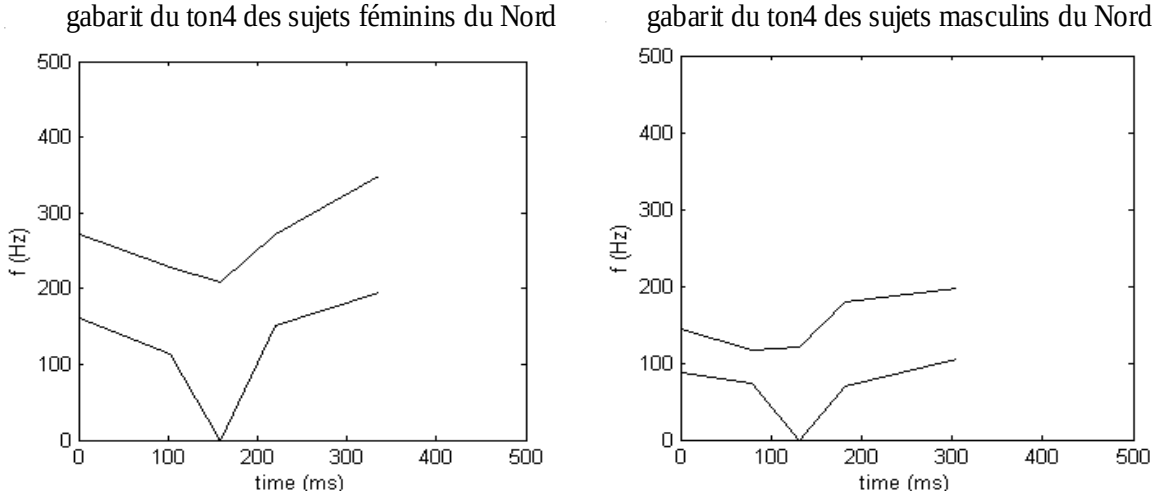


Figure 4.12 Gabarits du ton4 des sujets du Sud

4.4.1.5 Ton5

Le mouvement est en général montant. Il y a deux représentations : le ton5a dans les syllabes ouvertes, et le ton5b dans les syllabes fermées. Nous rappelons que la syllabe fermée est une syllabe qui se termine par l'une des trois consonnes /p/, /t/, /k/. Les syllabes restantes sont appelées les syllabes ouvertes.

Ton5a:

Le contour du ton5a contient deux segments. Le mouvement du pitch est horizontal ou descendant faiblement. Ensuite le mouvement est montant. Les points à mesurer sont donc le point initial du ton B, le point M où la direction du ton est changée et le point final E. La figure **4.13** donne un exemple du ton5a avec les points à mesurer.

Le tableau **4.14** et la figure **4.15** présentent les caractéristiques et les gabarits du ton5a. Le point initial est le même que pour le ton2. Le point final est plus haut que celui du ton1 et approximativement égal à celui du ton3.

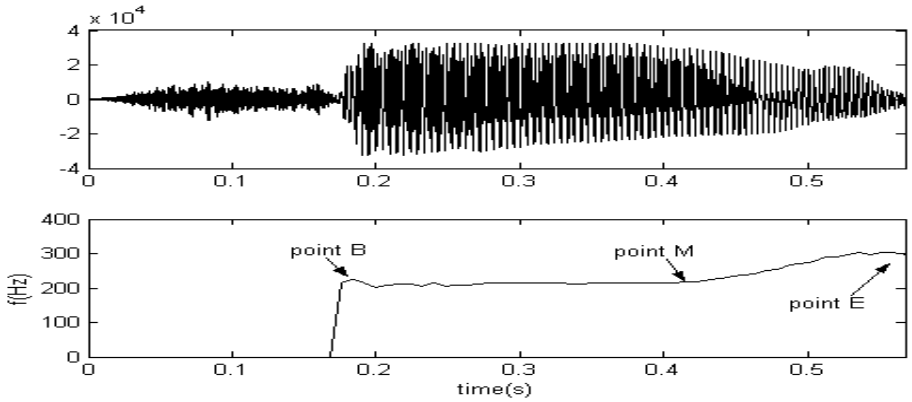


Figure 4.13 Un exemple du ton5a du sujet féminin VTT avec la syllabe "số"

Ton5b:

Le mouvement du pitch est monotone et montant. Les points à mesurer sont le point initial B et le point final E. La figure **4.14** donne un exemple du ton5b avec les points à mesurer.

Le tableau **4.15** et la figure **4.16** présentent les caractéristiques et les gabarits du ton5b. Le point initial du ton5b est plus haut que celui du ton5a et du ton1. La fréquence du point final

est approximativement la même que celle du ton5a et du ton3. La durée est plus courte que celle du ton5a.

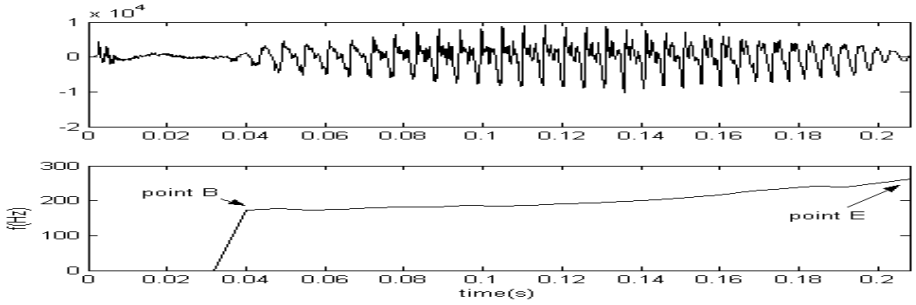


Figure 4.14 Un exemple du ton5b du sujet masculin TTA avec la syllabe "thoát"

locuteur	**fréquence** (Hz)														**durée** (ms)			
	point B				**point M**				**point E**				**delta**		**B → M**		**B → E**	
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type	moyenne	écart type
féminin			205 → 271	172 → 194			202 → 281				320 → 421	190 → 286						
PNY	216	11	246	181	209	7	225	188	290	25	355	231	74	26	235	32	392	246
VTT	201	13	250	172	198	9	222	175	272	21	333	219	70	22	201	46	351	44
DPQ	179	12	205	155	175	9	202	153	232	33	320	190	53	32	182	35	322	38
DHH	210	15	242	177	199	12	253	175	328	42	410	246	118	43	223	34	409	47
DHL	215	12	253	190	206	9	235	186	301	31	380	250	86	31	172	23	301	31
NTH	211	17	271	174	193	17	258	133	261	23	364	225	50	24	142	33	331	44
VTH	224	15	262	193	237	17	281	202	343	32	410	286	119	34	138	45	341	61
BKH	229	12	254	194	245	13	273	199	338	25	421	286	109	25	176	33	320	34
LPL	223	13	250	190	231	14	273	198	314	20	390	271	90	21	146	32	311	38
masculin			108 → 205	84 111			107 → 211	87 → 124			152 → 276	113 → 163						
BXH	116	6	128	101	116	8	145	100	167	14	197	130	51	13	168	25	334	38
TTA	139	10	166	110	143	10	167	124	205	22	276	163	66	20	161	42	302	51
LVS	127	11	205	109	133	14	211	107	202	28	266	158	74	28	117	42	324	54
TVH	97	4	108	84	96	3	107	87	128	7	152	113	30	6	142	25	275	34
HBQ	126	7	161	111	126	7	146	111	174	9	192	150	47	9	119	30	284	34
TTT	124	6	141	106	125	7	149	113	184	13	235	155	60	13	81	32	309	41

Tableau 4.14 Les caractéristiques acoustiques du ton5a

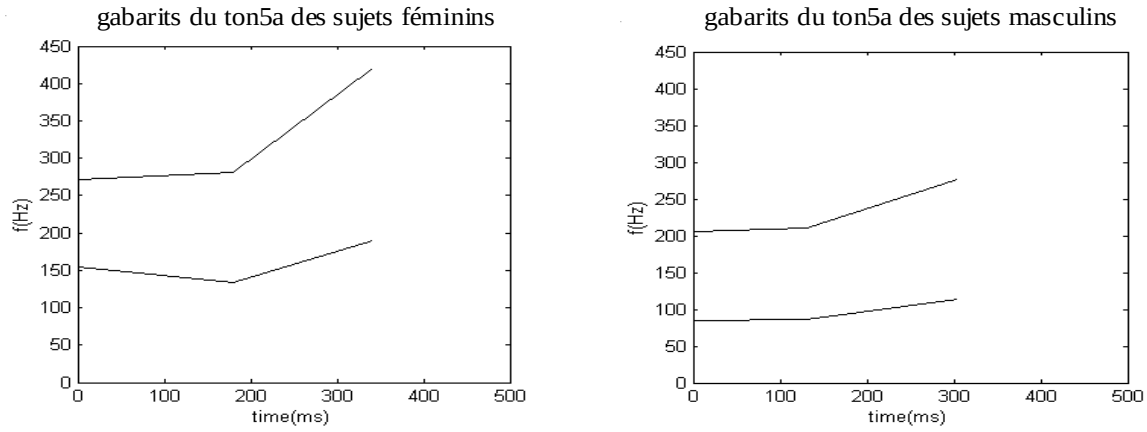


Figure 4.15 Gabarits du ton5a

locuteur	fréquence (Hz)										durée	
	point initial				point final				delta			
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type
féminin			253 →340	186 →238			326 →400	211 →296			99 →143	
PNY	285	23	340	238	335	21	372	296	50	25	111	44
VTT	251	18	285	213	296	20	346	266	45	23	129	46
DPQ	224	17	253	186	273	20	326	238	49	22	129	44
DHH	265	19	307	225	327	21	372	290	61	22	143	47
DHL	272	22	307	200	302	21	340	258	30	20	99	48
NTH	254	26	302	198	268	28	308	211	14	11	115	40
VTH	255	22	308	219	315	32	400	267	60	29	140	37
BKH	262	21	324	230	328	23	367	296	66	25	134	55
LPL	252	15	276	211	312	17	356	271	59	17	123	41
masculin			115 →219	90 →144			129 →246	110 →170			90 →134	
BXH	130	9	148	111	157	9	179	137	27	9	128	34
TTA	173	19	219	144	208	19	246	170	65	19	90	35
LVS	134	21	158	117	156	13	188	137	22	13	134	36
TVH	102	5	115	90	118	4	129	110	16	7	125	41
HBQ	142	14	168	122	171	11	190	152	28	13	129	44
TTT	133	9	160	109	158	10	179	140	24	11	111	35

Tableau 4.15 Les caractéristiques acoustiques du ton5b

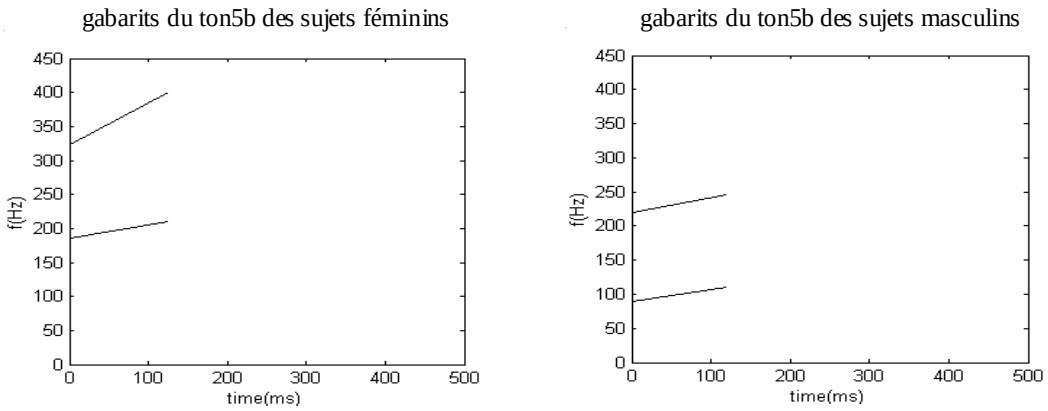


Figure 4.16 Gabarits du ton5b

4.4.1.6 Ton6

Le mouvement du ton6 est descendant en général. Il y a deux représentations : le ton6a dans les syllabes ouvertes, et le ton6b dans les syllabes fermées.

Ton6a:

Pour le ton6a nous avons trouvé quelques représentations différentes. Mais la plupart des sujets prononcent le ton6a horizontal au début puis descendant très fortement. Celui-ci est en accord avec les recherches précédentes. Les points à mesurer sont le point initial B, le point M où la direction du contour est changée et le point final E. La figure **4.17** donne un exemple du ton6a avec les points à mesurer.

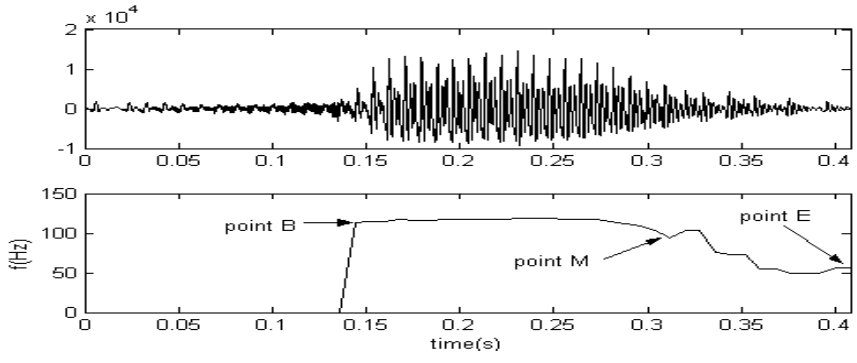


Figure 4.17 Un exemple du ton6a du sujet masculin BXH du Nord avec la syllabe "dụ"

Par ailleurs nous avons trouvé quelques autres contours du ton6a de la voix sud et centrale.

Pour le sujet féminin BKH du Sud le contour du ton6a présente une rupture au milieu. Le figure **4.18** présente un exemple. Dans ce cas, le contour est semblable aux contours de ton3 et ton4. Néanmoins la pente n'est pas grande comme pour le ton3 ou le ton4.

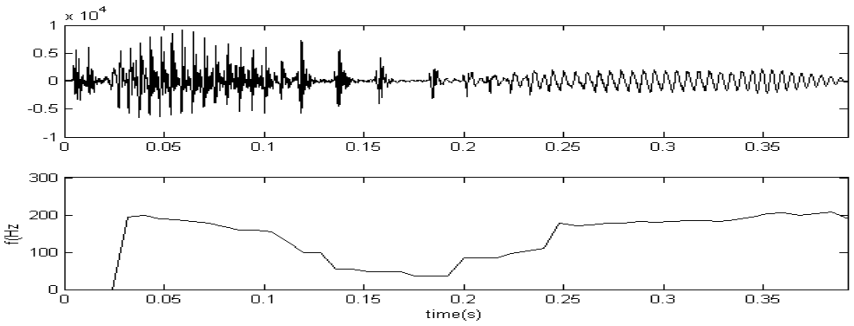


Figure 4.18 Un exemple du ton6a du sujet féminin BKH du Sud (la syllabe "cộng") dont contour a la rupture au milieu

Pour le sujet féminin NTH du Centre, le contour du ton6a est descendant et montant. La partie descendante est plus longue que la partie montante. La figure **4.19** donne un exemple du ton6a du sujet féminin NTH.

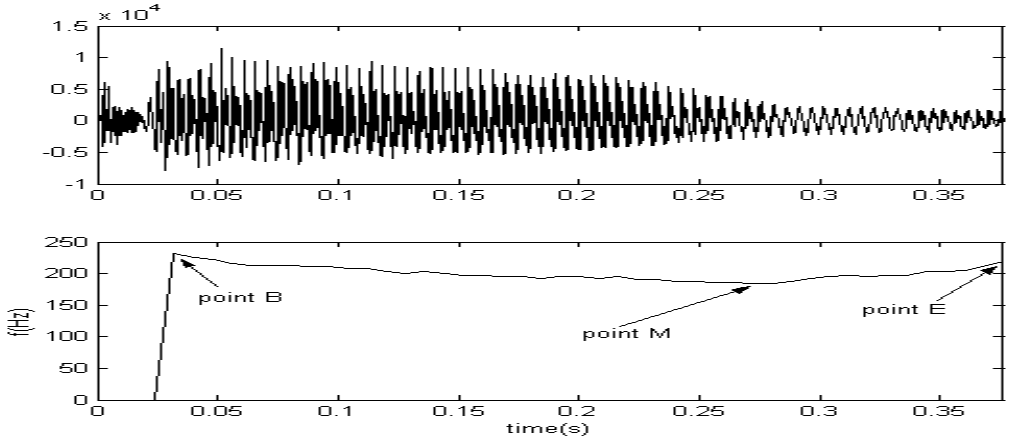


Figure 4.19 Un exemple du ton6a du sujet féminin NTH du Centre avec la syllabe "cạnh"

Pour le cas du sujet masculin TTT du Sud, le ton6a contient deux segments descendant et montant également. La figure **4.20** présente un exemple du ton6a du sujet TTT.

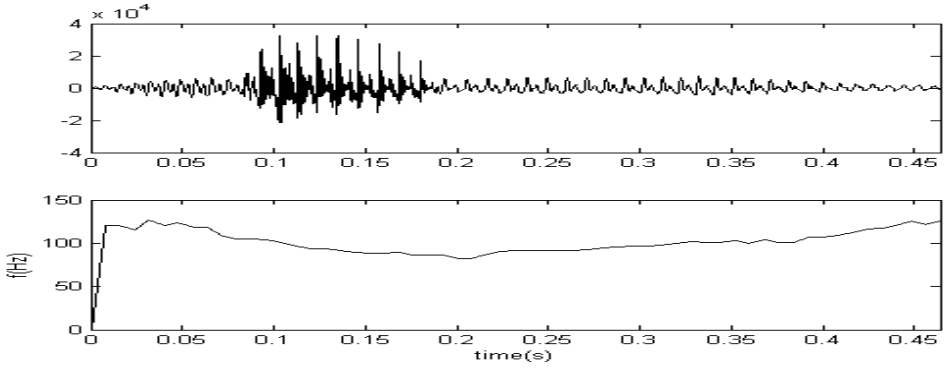


Figure 4.20 Un exemple du ton6a du sujet masculin TTT du Sud avec la syllabe "bịa"

En cas du sujet masculin HBQ du Sud, du sujet masculin TVH du Centre, du sujet masculin LVS du Centre, le ton6a varie. Le contour de ton6a peut être descendant ou descendant-montant. La figure **4.21** donne les 2 représentations différentes du ton6a du sujet LVS pour la même syllabe "sự".

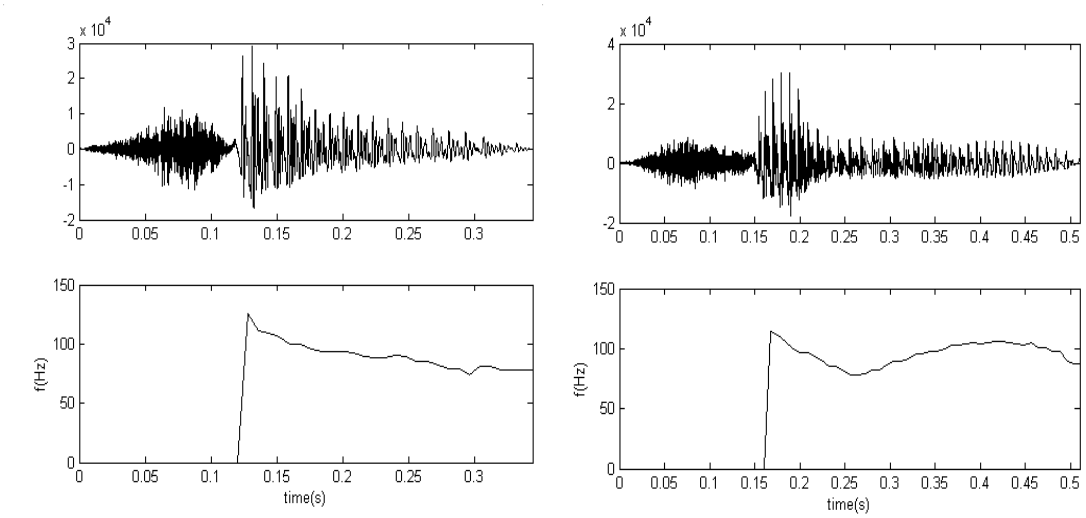


Figure 4.21 Un exemple des deux contours du ton6a du sujet LVS du Centre avec la même syllabe "sự"

Le tableau **4.16** et la figure **4.22** présentent les caractéristiques et les gabarits généraux du ton6a. Le point initial est au même niveau que pour le ton2 et le ton5a. Le point final est très bas.

locuteur	fréquence (Hz)														durée (ms)			
	point B				point M				point E				delta		B → M		B → E	
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type	moyenne	écart type
féminin			213 →275	144 →184			213 →275				163 →235	0 →48					103 →223	
PNY	215	19	246	183	209	20	242	179	98	47	219	44	-117	48	90	41	140	44
VTT	204	14	242	175	213	18	246	122	116	49	235	48	-88	51	108	29	159	45
DPQ	176	15	213	144	183	14	216	150	94	38	179	30	-82	40	90	26	153	37
DHH	207	12	239	184	195	12	235	176	74	36	184	33	-133	40	123	36	187	40
DHL	213	19	275	175	217	17	271	177	159	52	228	43	-54	55	67	26	103	23
VTH	217	24	271	178	186	10	213	102	72	27	163	0	-145	40	118	36	223	51
LPL	202	13	234	178	179	10	208	155	47	45	188	0	-154	45	103	42	177	35
masculin																		
BXH	109	9	126	87	108	9	126	84	44	12	83	23	-65	16	75	33	186	45
TTA	134	7	149	111	136	9	164	108	83	25	135	50	-51	25	75	22	125	27

Tableau 4.16 Les caractéristiques acoustiques du ton6a

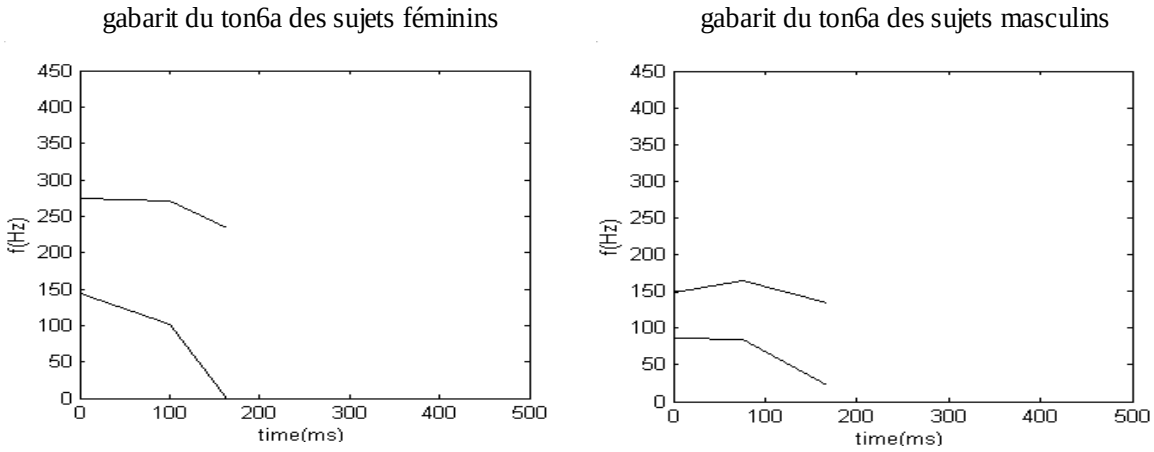


Figure 4.22 Gabarits du ton6a

Ton6b:

Le contour est descendant et monotone. Les points à mesurer sont donc le point initial B et le point final E. La figure **4.23** donne un exemple du ton6b avec les points à mesurer.

Le tableau **4.17** et la figure **4.24** présentent les caractéristiques et les gabarits du ton6b. Le point initial est au même niveau que celui du ton6a et du ton2. Le point final est plus haut que celui du ton2. La durée est très courte.

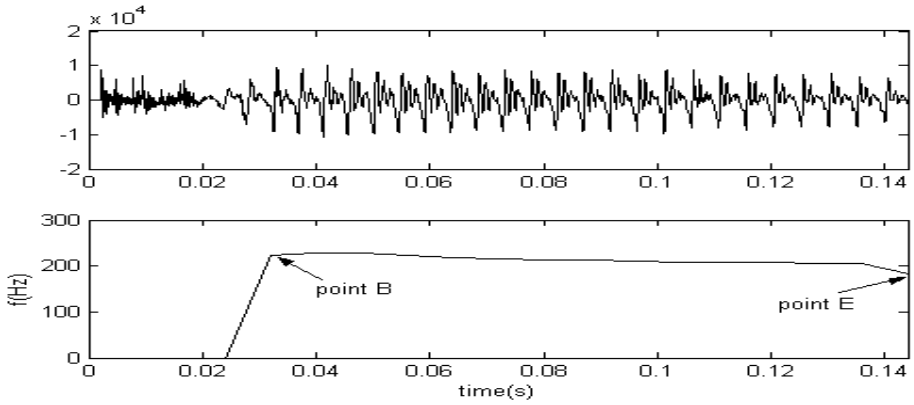


Figure 4.23 Un exemple du ton6b du sujet féminin LPL avec la syllabe "cặp"

locuteur	fréquence (Hz)										durée	
	point initial				point final				delta			
	moyenne	écart type	max	min	moyenne	écart type	max	min	moyenne	écart type	moyenne	écart type
féminin			205 →264	166 →208			195 →219	100 →186			93 →130	
PNY	216	8	231	205	188	25	216	100	-28	23	93	32
VTT	207	7	219	186	189	11	207	158	-17	12	129	46
DPQ	185	9	205	166	164	10	195	148	-21	12	129	44
DHH	214	9	235	195	188	9	216	170	-26	7	122	32
DHL	214	9	231	200	193	9	205	175	-21	8	99	48
NTH	230	13	262	208	193	8	219	186	-32	14	126	37
VTH	223	21	264	195	189	23	211	180	-30	25	130	31
BKH	218	12	242	188	184	12	205	160	-34	18	119	39
LPL	211	8	228	197	183	7	198	167	-28	9	129	27
masculin			115 →146	89 →125			100 →137	72 →110			87 →116	
BXH	112	5	125	105	105	5	116	99	-6	4	116	24
TTA	134	6	146	125	119	6	137	110	-14	4	87	18
LVS	107	5	115	98	89	7	103	72	-18	7	103	32
TVH	99	6	121	89	91	4	100	80	-8	7	105	30
HBQ	121	5	134	111	110	5	122	100	-10	3	109	35
TTT	111	7	128	96	94	5	108	85	-16	10	99	32

Tableau 4.17 Caractéristiques acoustiques du ton6b

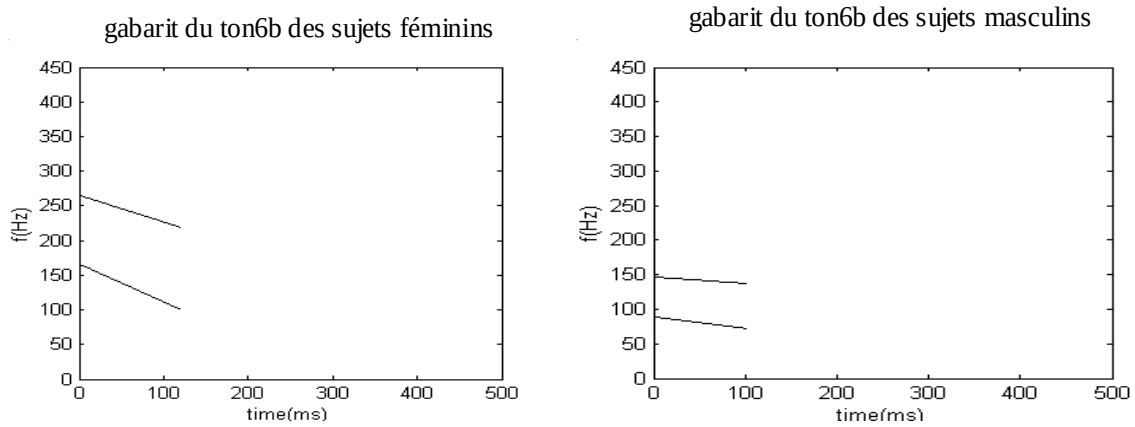


Figure 4.24 Gabarits du ton6b

4.4.2 Conclusion sur les six tons vietnamiens

4.4.2.1 Contour des tons

Nous avons quelques commentaires sur les contours des tons:

- Le ton1 est monotone. Le contour du ton1 est horizontal et plat. Les points initial et final se terminent en haute fréquence.
- Le ton2 est monotone. Le contour du ton2 est descendant et plat. Le point initial est plus bas que celui du ton1.
- Le contour du ton3 est complexe. Il est descendant d'abord puis montant. Le ton3 peut avoir une rupture ou non au milieu. Le point initial du ton3 est plus bas que celui du ton1. Le point final du ton3 est plus haut que celui du ton1.
- Le ton4 a deux représentations: le ton4 des sujets du Nord et le ton4 des sujets du Sud et du Centre.
  - Pour le ton4 des sujets du Nord d'après Andrejev et Gordina [Andreev et Gordina, 1957], le ton4 possède le contour qui est descendant et puis montant. La partie descendante est proportionnée avec la partie montante. D'après Vu [Vu, 1984], le ton4 est descendant en général jusqu'au 2/3 du ton et puis il monte faiblement. D'après Vu [Vu, 1999], pour la voix féminine, le ton4 est descendant jusqu'au 3/4 du ton et montant faiblement, alors que pour la voix masculine, le ton4 est descendant puis montant, les deux parties étaient proportionnées. D'après nos résultats, le ton4 est descendant jusqu'au 4/5 du ton puis horizontal ou montant faible. Ceci peut être expliqué par le fait que la prononciation du ton4 est assez difficile lorsque le contour est descendant et montant également. Actuellement les jeunes ont tendance à simplifier la façon de prononcer le ton4. Le contour du ton4 possède donc deux parties. La première partie du contour est descendante. La deuxième partie du contour est horizontale ou monte faiblement et elle est plus courte que la première partie.
  - Le ton4 des sujets du Sud et du Centre est prononcé sans distinction avec le ton3.
- Le contour du ton5 est montant en général. Il y a deux représentations : le ton5a dans la syllabe ouverte et le ton5b dans la syllabe fermée. Le contour du ton5a est horizontal au début et montant ensuite. Le contour du ton5b est montant avec une durée courte. Le point initial du ton5b est plus haut que celui du ton5a.

- Le ton6a a deux représentations comme le ton5, c'est-à-dire le ton6a dans la syllabe ouverte et le ton6b dans la syllabe fermée.
  - Le contour du ton6a est en général horizontal et ensuite descendant très rapidement. Le point initial est au même niveau que celui du ton2 et du ton5a. Le point final est très bas. Il y a quelques représentations spéciales du ton6a pour quelques sujets du Sud et Centre. Par exemple le contour est descendant et montant comme le cas du sujet NTH du Centre. Dans le cas du sujet BKH du Nord, le contour présente une rupture au milieu, etc. Pour ces contours, on aura besoin d'un ensemble plus important de sujets avant d'arriver à une conclusion exacte.
  - Le contour du ton6b est descendant. Le point initial est au même niveau que celui du ton2. La durée est très courte.

4.4.2.2 Registre des tons

Le registre est la hauteur relative de ton fondée sur l'impression auditive. Six tons sont classifiés en deux registres : le registre haut et le registre bas. [Adreev et Gordina, 1957] ont utilisé la hauteur du point final comme le critère de classification des registres. Pour ces auteurs, le point final a une signification plus importante. Les résultats sont en accord avec les autres : les ton1, ton3, ton5 sont au registre haut et les ton2, ton4 et ton6 sont au registre bas. Ceci est tout à fait en accord avec nos résultats : le point final du ton1, du ton3 et du ton5 est plus haut que celui du ton2, du ton4 et du ton6.

4.4.2.3 Durée des tons

La durée des tons dépend de l'articulation de chaque personne, donc pour avoir une mesure normalisée de la longueur d'un ton, nous avons utilisé la durée relative [Vu, 1999] :

$$X = \frac{Y}{N}$$

Où N est la durée moyenne des six tons, Y est la durée moyenne du ton i et X est la durée relative du ton i. Les figures **4.25** et **4.26** présentent les diagrammes de la durée relative des six tons des sujets du Nord, du Sud et du Centre.

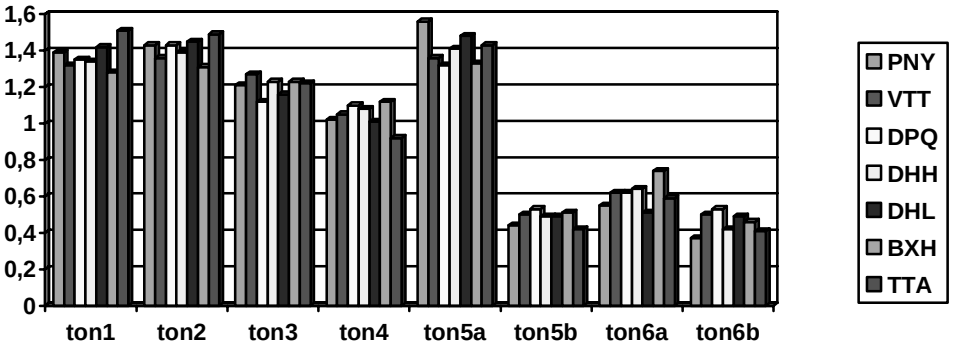


Figure 4.25 Diagramme de la durée relative des 6 tons des sujets du Nord

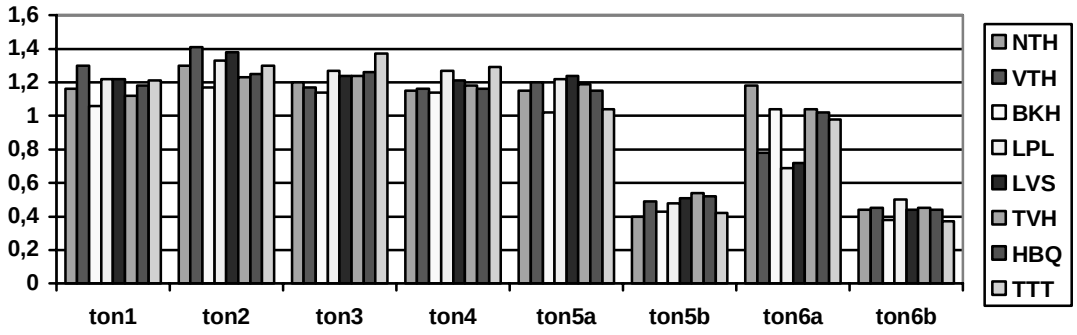


Figure 4.26 Diagramme de la durée relative des 6 tons des sujets du Sud

Le ton, dont la durée relative est supérieure à 1, est considéré comme un ton long. Dans le cas contraire, il est considéré comme un ton court. Nous avons:

- Le ton1, le ton2, le ton3 et le ton5a qui sont les tons longs.
- Le ton5b et le ton6b qui sont les tons courts.
- Le ton4 des sujets du Sud et du Centre est un ton long. Pour le ton4 des sujets du Nord, il est prononcé avec une durée relative supérieure à 1 par la plupart des sujets. En général, nous considérons que le ton4 du Nord est aussi un ton long.
- Le ton6a des sujets du Nord est un ton court. Pour les sujets du Sud et du Centre, la durée du ton6a est largement variable. Ceci dépend de la prononciation particulière de chaque sujet.

4.5 Conclusions

Dans ce chapitre nous avons évalué notre méthode de détermination de pitch, fondée sur l'analyse en ondelettes. Nous avons utilisé le corpus d'un homme et d'une femme de [Bagshaw et al 1993] et comparé avec le contour de pitch de référence déterminé à partir de laryngographe. Les résultats ont montré que la méthode est performante pour déterminer le pitch et ils sont comparables avec les autres méthodes.

Le vietnamien est une langue à tons. Dans ce chapitre nous avons présenté les caractéristiques et les gabarits généraux des six tons vietnamiens. Pour les sujets du Nord, les représentations des six tons sont assez identiques parmi les sujets. Pour les sujets du Sud et du Centre, le ton3 et le ton4 sont prononcés identiquement. Nous avons trouvé quelques représentations particulières du ton6a des sujets du Sud et du Centre. Ceux-ci peuvent être des représentations des dialectes ou seulement une façon spéciale de prononcer pour certains sujets. On a besoin d'examiner sérieusement ce cas avant de donner une conclusion. Cette caractéristique peut avoir une influence sur la reconnaissance des tons que nous présenterons plus tard. Nous ferons donc la reconnaissance des tons sur les voix du Nord.

4.6 Références

Andreev N.D et Gordina M.V (1957)
Н. Д. Андреев et М. В. Гордина
"Système des tons vietnamiens" (en russe)
по эспериментальым данным, Весник, No 8

Bagshaw P.C, Hiller S.M, Jack M.A (1993)
*Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. (*http://www.cstr.ed.ac.uk/~pcb/fda_eval.tar.gz*)*
Proceedings of EuroSpeech'93, Berlin.

Doan T.T (1977).
Ngữ âm tiếng Việt *(Phonétique vietnamienne)*
Nhà xuất Đại học và Trung học chuyên nghiệp

Gold. B et Rabiner. L (1969)
Parallel processing techniques for estimating pitch periods of speech in the time domain
Acoustical Society of America, 46(2), pp442-448.

Han M. S, Kim K. O (1974).
Phonetic variation of Vietnamese tones in disyllabic utterances tones
Journal of Phonetics, vol. 2, pp 223-232

Hess. W.H (1983)
Pitch Determination of Speech Signal: Algorithms and Devices
Springer-Verlag, Heidelberg, Germany.

Hoang T, Hoang M (1975)
Remarques sur la structure phonologique du vietnamien
Etudes des vietnamiens, No 40, Hanoi

Kadambe S, Boudreaux-Bartels G.F (1992)
Application of the Wavelet Transform for Pitch Detection of Speech Signals
IEEE Trans. Information Theory, vol. 38, no 2, March.

Kadambe S, Boudreaux-Bartels G.F. (1991)
A Comparison a Wavelet Functions for Pitch Detection of Speech Signals
In Proc. IEEE ICASSP 1991

Mallat S. G. et S. Zhong (1989)
Complete signal representation with multi scale edges
Tech. rep RRT-483-RR-219, Courant Inst. of Math. Sci., Dec. 1989.

Medan Y, Yair. E et Chazan D (1991)
Super resolution pitch determination of speech signals
IEEE Trans. Signal Processing, ASSP vol 39(1), pp 40-48.

Noll A.M (1967)
Cepstrum pitch determination
Journal of the Acoustical Society of America, vol 41(2), pp 239-309

Noll A.M (1970)
Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate
Symposium on Computer Processing in Communication, vol 19, pp 779-797. Poly-technique Institute of Brooklyn Microwave Research Institute, New York

Phillips M.S (1985)
A feature-based time domain pitch tracker
Journal of the Acoustical Society of America, vol 77: S9-S10(A)

Rabiner L,.R, Cheng M.J, Rosenberg A.E and McGonegal C.A (1976)
A Comparative performance study of several pitch detection algorithms.
IEEE Trans. Audio, Signal and Speech Processing, vol 24, pp 399-417

Schroeder M.R (1968)
Period histogram and product spectrum: New methods for fundamental frequency measurement
Journal of the Acoustical Society of America, vol 43(4), pp 829-834

Secrest B.G et Doddington (1983)
An integrated pitch tracking algorithm for speech systems
Proc. IEEE ICASSP-83, pp 1352-1355, Boston

Vu B. H (1999).
"Les caractéristiques fondamentales des tons vietnamiens dans leur état statique" (en vietnamien)
Ngôn ngữ, Vol. 6, pp 34-53

Vu K. B (1984)
"Untersuchungen zu den wesentlichen akustischen parametern der vietnamesischen silben (Grundfrequenz - Intensitätsverlauf und Dauer)
"Études aux paramètres acoustiques essentiels des syllabes vietnamiennes (fréquence de base - cours d'intensité et durée)
Phill. Diss, HU, Berlin, 1984

Wendt C, Petropulu A.P (1996)
Pitch determination and speech segmentation using the discrete wavelet transform
IEEE International Symposium on Circuits and Systems, vol 2, pp 45-48.

5

Reconnaissance des tons vietnamiens en mode mots isolés

La reconnaissance des tons est une tâche importante dans un système de reconnaissance de la parole pour les langues à tons. Comme nous l'avons vu dans le chapitre précédant les tons sont caractérisés par le registre, le contour de pitch et la durée. Le vecteur acoustique contiendra donc ces éléments, l'énergie sera également utilisée. La reconnaissance des tons peut être réalisée avec les techniques de réseau de neurones [Lee et al 1995] ou les techniques de chaînes de Markov caché ou HMM [Yang et all, 1988]. Pour la reconnaissance des tons vietnamiens nous préférons utiliser les techniques HMM. Le contour de pitch est extrait à partir de l'algorithme ondelette [Kadambe & Boudreaux-Bartels, 1992][Wendt & Petropulu, 1996] que nous avons implanté et décrit dans le chapitre 4. La reconnaissance est réalisée sur les six tons standards des sujets du Nord.

5.1 Reconnaissance des tons vietnamiens

Pour la reconnaissance des tons vietnamiens nous avons choisi d'utiliser une méthode fondée sur les HMMs (implémentée avec les outils de HTK [Young et al, 2000]) pour caractériser la variation temporelle des contours de pitch. Chaque ton est modélisé par un HMM : pour les six tons vietnamiens, nous utilisons donc 8 HMMs: quatre HMMs pour le ton1, 2, 3 et 4 plus quatre HMMs pour le tons5a, le tons5b, le ton6a et le tons6b. Nous utilisons des HMMs avec 3 états émettant des observations. Une gaussienne est utilisée pour chaque état.

L'implémentation des HMMs est réalisée en deux phases: la phase d'apprentissage et la phase de classification. Pendant la phase d'apprentissage les HMMs sont entraînés à partir des données de l'ensemble d'apprentissage.

L'information du ton est présente sur la partie finale de chaque syllabe. Néanmoins pour la reconnaissance automatique du ton, la détection de la partie finale est assez difficile lorsque la partie initiale est voisée car la frontière entre les deux n'est pas nette. Cette difficulté est de même nature que celle qui apparaît lors de la détection de la frontière entre une consonne voisée et une voyelle. C'est pour cette raison que nous avons choisi de ne reconnaître le contour de pitch uniquement que sur la partie voisée de la syllabe. Le schéma du système de reconnaissance des tons est présenté en figure **5.1**.

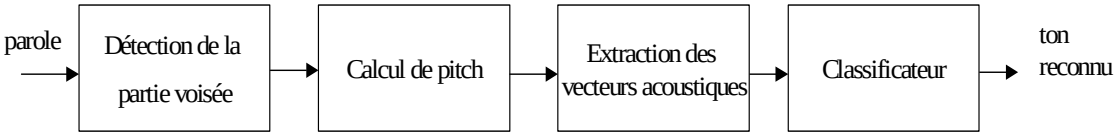


Figure 5.1 Principe du système de reconnaissance des tons

5.1.1 Vecteur caractéristique du Mandarin

Pour la reconnaissance des 4 tons du Mandarin, [Yang et al 1988] ont utilisé un vecteur caractéristique qui est défini comme suit:

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1})] \qquad (5.1)$$

où f_t est la fréquence du pitch de la trame *t*.

La première composante est proportionnelle à la pente locale de logarithme du pitch, alors que la deuxième est proportionnelle au double de la valeur logarithmique de l'amplitude local. Dans la même étude pour éliminer les variations indésirables ainsi que pour la reconnaissance des tons indépendante du locuteur, le pitch est normalisé par une valeur de référence du pitch F_0 qui dépend du locuteur. Chaque valeur f_t du vecteur caractéristique est alors remplacée par le rapport (f_t/F_0). Seule la deuxième composante du vecteur est affectée par F_0. Pour la reconnaissance des 4 tons lexicaux du Mandarin le taux de reconnaissance obtenu est ~97%.

Nous décidons pour un premier essai d'utiliser ce vecteur caractéristique du Mandarin pour l'appliquer à la reconnaissance des tons vietnamiens. La langue vietnamienne présente deux

tons monotones, le ton1 dans le registre haut et le ton2 dans le registre bas. Pour normaliser les valeurs du pitch d'un locuteur, la valeur de référence F_0 doit donc être choisie à partir de la valeur moyenne du ton1 ou du ton2 pour un même locuteur dans nos expériences.

5.1.2 Expérimentation

Le contour de six tons dépend du dialecte. En général nous classifions trois dialectes : Nord, Centre et Sud. Pour notre étude de la reconnaissance des tons du vietnamien standard (dialecte Hanoi), nous avons extrait de notre corpus les signaux de parole prononcés par les 5 sujets féminins de Hanoi.

Les buts de nos expériences sont de savoir:

- Quel est l'effet de la voyelle pour la reconnaissance des tons? Dans ce cas les données d'entraînement contiendront seulement les syllabes liées à 8 voyelles parmi les 16. Les données de tests sont différentes des données d'entraînement et séparées en 2 tests. Le premier test (test1) contiendra les syllabes liées aux 8 voyelles existant dans les données d'entraînement. Le deuxième test (test2) contiendra les syllabes liées aux 8 autres voyelles qui ne sont pas représentées dans les données d'entraînement.
- Quel est l'effet de la normalisation des tons pour la reconnaissance des tons vietnamiens? Pour la normalisation, les monotones, (ton1 et ton2), seront utilisés. C'est à dire que la valeur moyenne du ton1 ou du ton2 d'un locuteur est utilisée pour normaliser toutes les valeurs du pitch pour le locuteur.
- Les expériences seront réalisées en deux modes: la reconnaissance dépendante du locuteur et la reconnaissance indépendante du locuteur:
  - Pour la reconnaissance dépendante du locuteur, les données des quatre sujets seront utilisées pour l'entraînement et pour les tests
  - Pour la reconnaissance indépendante du locuteur, les données du cinquième locuteur seront utilisées pour les tests

  On utilise toutes les permutations pour 5 locuteurs disponibles. Le résultat global est obtenu en effectuant la moyenne de résultats de cinq expériences. Chaque fois les données d'entraînement contiennent 1051 syllabes. Pour les tests dépendants du locuteur, le test1 contient 361 syllabes, le test2 contient 640 syllabes. Pour les tests indépendants du

locuteur, le test1 contient 354 syllabes et le test2 contient 160 syllabes. Ces conditions sont résumées dans la figure **5.2**. Bien sûr, nous précisons quand même que dans tous les cas, les données d'apprentissage et de test sont disjointes.

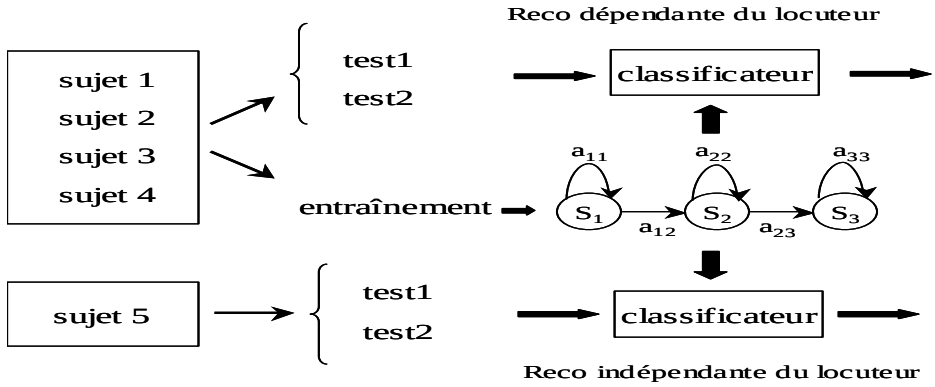


Figure 5.2 Résumé des conditions expérimentales

Les 3 premières expériences sont ainsi menées à la fois pour valider l'utilisation du vecteur caractéristique proposé pour le Mandarin et pour déterminer la meilleure valeur de référence F_0 (calculée à partir du ton1 ou du ton2) :

- Expérience 1 : utilisation du vecteur caractéristique du Mandarin sans normalisation tel qu'il est défini par l'équation (5.1).
- Expérience 2 : utilisation du vecteur défini par l'équation (5.1) mais normalisation par une valeur de référence de F0 calculée à partir de la moyenne du ton1.
- Expérience 3 : même calcul que l'expérience précédente mais avec une normalisation par rapport au ton monotone ton2.

Pour modéliser les tons, des HMMs avec 3 états sont utilisés.

Les résultats sont présentés dans le tableau **5.1**. Pour chaque tableau, la première colonne présente les expérimentations. Nous avons 5 sujets notés {VTT, PNY, DPQ, DHL, DHH}. L'expérimentation VTT signifie que les données des quatre sujets {PNY, DPQ, DHL, DHH}sont utilisées pour l'entraînement et les tests dépendants du locuteur, et que les données du sujet VTT sont utilisées pour les tests indépendants du locuteur. Les taux de reconnaissance de chaque expérimentation présentés sont calculés de manière simple par le rapport entre le nombre de tons reconnus sur le nombre total de tons du test.

L'utilisation du vecteur défini dans l'équation (5.1) donne des taux de reconnaissance (expérience 1) toujours inférieurs de 5% aux mêmes tests réalisés en normalisant par la fréquence fondamentale (expériences 2 et 3) en mode de reconnaissance dépendante du locuteur. L'effet de la normalisation est encore plus évident en mode de reconnaissance indépendante du locuteur. L'utilisation d'une fréquence de référence semble donc améliorer les résultats. Néanmoins le calcul de cette moyenne à partir du ton1 ou du ton2 donne des résultats similaires.

	taux de Reco (%)	
	test1	test2
VTT	82.5	86.1
PNY	77.8	84.3
DPQ	88.3	90.2
DHL	78.1	81.4
DHH	79.3	82.1
moyenne	81.2	84.8

(a)

	taux de Reco (%)	
	test1	test2
VTT	91.7	93.6
PNY	86.7	89.9
DPQ	84.7	88.8
DHL	87.2	88.9
DHH	85.1	87.9
moyenne	87.1	89.8

(b)

	taux de Reco (%)	
	test1	test2
VTT	90.3	93.1
PNY	86.4	89.9
DPQ	85.3	88.5
DHL	86.7	88.9
DHH	86.2	88.6
moyenne	87.0	89.8

(c)

	taux de Reco (%)	
	test1	test2
VTT	75.0	77.2
PNY	80.8	83.8
DPQ	41.0	35.0
DHL	87.6	93.2
DHH	85.7	89.3
moyenne	74.0	75.7

(d)

	taux de Reco (%)	
	test1	test2
VTT	75.3	77.2
PNY	88.7	88.1
DPQ	89.0	90.0
DHL	89.5	93.8
DHH	92.6	90.6
moyenne	87.0	87.9

(e)

	taux de Reco (%)	
	test1	test2
VTT	75.6	78.4
PNY	86.2	88.1
DPQ	87.3	88.8
DHL	89.0	91.9
DHH	90.0	86.2
moyenne	85.6	86.7

(f)

Tableau 5.1. Taux de reconnaissance des six tons vietnamiens utilisant le vecteur avec 2 composantes. Les trois premiers tableaux présentent les résultats dans le mode dépendant du locuteur. Les trois derniers présentent les résultats dans le mode indépendant du locuteur: a, d) sans normalisation b, e) normalisé par le ton1 c, f) normalisé par le ton2.

Toutefois, la performance maximale de nos essais atteint seulement ~88.4% en mode de reconnaissance dépendant du locuteur et ~87.5% en mode de reconnaissance indépendant du locuteur alors que les résultats proposés par [Yang et al, 1988] pour les quatre tons lexicaux du Mandarin montrent un maximum de 98.3% en mode dépendant du locuteur et de 96.5% en mode indépendant du locuteur. De plus, en première approximation, les résultats du 1er test réalisé avec les mêmes voyelles que la phase d'apprentissage sont moins bons que ceux du 2ème test avec des voyelles différentes (c'est-à-dire des conditions différentes). Ceci montre

que notre reconnaissance de tons est relativement indépendante de la voyelle considérée. Nous pouvons déjà conclure que le vecteur utilisé pour le Mandarin n'est pas optimal pour le vietnamien. Nous pouvons consolider cette hypothèse en remarquant que les contours de pitch des quatre tons chinois semblent plus simples que les six tons vietnamiens avec 8 représentations.

5.1.3 Vecteur caractéristique adapté au vietnamien

Nous nous sommes inspiré des idées de [Lee L-S et al, 1993] qui a utilisé les informations supplémentaires de la durée et de l'énergie pour la reconnaissance des quatre tons lexicaux et d'un ton neutre du Mandarin, et de [Lee T et al, 1995] qui a utilisé aussi les informations supplémentaires de la durée et l'énergie pour la reconnaissance des neuf tons lexicaux du Cantonais. Si on regarde les particularités des tons vietnamiens, on remarque:

- La différence du contour de pitch entre le point à la fin et le point au début, est importante. Si la différence est positive, alors le ton possible est entre le ton1, le ton3, le ton5a et le ton5b. Si la différence est négative alors le ton possible est entre le ton1, le ton2, le ton4, le ton6a ou le ton6b.
- Si la durée du ton est courte, alors il est possible d'avoir le ton5b, ton6a ton6b ou ton4.

Les informations supplémentaires sur la différence entre le point à la fin et le point au début, de la durée et de l'énergie peuvent améliorer la reconnaissance des tons vietnamiens. Celles-ci seront calculées de la façon suivante:

- $log(f_e)$ - $log(f_m)$: la partie initiale de la syllabe est influencée par la consonne voisée si celle-ci existe dans la syllabe. La différence est donc calculée entre le point milieu (f_m) et le point terminal (f_e) de la syllabe.
- a : la durée de la partie voisée.
- $log(e_t)$ - $log(e)$: la normalisation d'énergie avec e_t l'énergie à la trame t, e l'énergie maximale de la partie voisée.

Nous combinons ces trois composantes au vecteur défini en (5.1), appelé *F*. Nous allons avoir alors 7 types de vecteurs. Ces sept vecteurs seront alors notés :

- Vecteur *Fe* : vecteur *F* ajouté à l'énergie.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), \log(e_t) - \log(e)] \quad (5.2)$$

- Vecteur *Ff* : vecteur *F* ajouté à la différence entre le point terminal et le point milieu.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), \log(f_e) - \log(f_m)] \quad (5.3)$$

- Vecteur *Ft* : vecteur *F* ajouté à la durée.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), a] \quad (5.4)$$

- Vecteur *Fef* : vecteur *F* ajouté à l'énergie et à la différence entre le point terminal et le point milieu.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), \log(e_t) - \log(e), \log(f_e) - \log(f_m)] \quad (5.5)$$

- Vecteur *Fet* : vecteur *F* ajouté à l'énergie et à la durée.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), \log(e_t) - \log(e), a] \quad (5.6)$$

- Vecteur *Ftf* : vecteur *F* ajouté à la durée et à la différence entre le point terminal et le point milieu.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), a, \log(f_e) - \log(f_m)] \quad (5.7)$$

- Vecteur *Fetf*: vecteur *F* ajouté à l'énergie, à la durée et à la différence entre le point terminal et le point milieu.

$$y_t = [\log(f_t) - \log(f_{t+1}), \log(f_t) + \log(f_{t+1}), \log(e_t) - \log(e), a, \log(f_e) - \log(f_m)] \quad (5.8)$$

Les expérimentations réalisées avec le vecteur *F* sont refaites avec ces sept vecteurs différents.

Les taux de reconnaissance en moyenne des huit vecteurs (5.1) - (5.8) sont présentés dans les tableaux **5.2**, **5.3**, **5.4**.

type du vecteur	taux de reco		
	test1	test2	moyenne
F	81.2	84.8	83.0
Fe	82.3	87.0	84.6
Ff	84.5	83.0	83.8
Ft	84.2	85.7	84.9
Fef	86.9	86.0	86.5
Fet	84.6	84.9	84.7
Ftf	90.2	90.8	90.5
Fetf	90.5	91.2	90.8

reconnaissance dépendante du locuteur

type du vecteur	taux de reco		
	test1	test2	moyenne
F	74.0	75.7	74.8
Fe	76.9	79.2	78.0
Ff	79.9	78.9	79.4
Ft	78.5	78.9	78.7
Fef	82.3	81.9	82.1
Fet	79.1	79.6	79.3
Ftf	85.8	83.9	84.8
Fetf	86.4	86.7	86.5

reconnaissance indépendante du locuteur

Tableau 5.2. Taux de reconnaissance utilisant les vecteurs sans normalisation

type du vecteur	taux de reco		
	test1	test2	moyenne
F	87.1	89.8	88.4
Fe	87.4	89.9	88.6
Ff	87.6	87.8	87.7
Ft	87.9	89.8	88.9
Fef	89.2	88.2	88.7
Fet	88.5	90.4	89.5
Ftf	93.1	92.7	92.9
Fetf	93.7	92.8	93.3

reconnaissance dépendante du locuteur

type du vecteur	taux de reco		
	test1	test2	moyenne
F	87.0	87.9	87.5
Fe	87.0	88.6	87.8
Ff	87.4	86.1	86.7
Ft	86.7	87.4	87.1
Fef	89.5	87.7	88.6
Fet	86.1	87.3	86.7
Ftf	91.2	91.8	91.5
Fetf	91.4	91.6	91.5

reconnaissance indépendante du locuteur

Tableau 5.3. Taux de reconnaissance avec les vecteurs normalisés par le ton1

type du vecteur	taux de reco		
	test1	test2	moyenne
F	87.0	89.8	88.4
Fe	87.4	90.5	89.0
Ff	88.1	87.3	87.7
Ft	87.6	89.5	88.6
Fef	89.6	88.2	88.9
Fet	88.4	90.3	89.3
Ftf	93.1	92.1	92.6
Fetf	93.7	92.6	93.2

reconnaissance dépendante du locuteur

type du vecteur	taux de reco		
	test1	test2	moyenne
F	85.6	86.7	86.1
Fe	87.0	89.2	88.1
Ff	87.0	86.3	86.6
Ft	85.9	87.0	86.5
Fef	89.2	87.3	88.2
Fet	86.3	87.9	87.1
Ftf	91.2	91.4	91.3
Fetf	91.2	91.7	91.4

reconnaissance indépendante du locuteur

Tableau 5.4. Taux de reconnaissance avec les vecteurs normalisés par le ton2

On peut faire les commentaires suivants sur les résultats:

- Les taux de reconnaissance du vecteur *Fetf* avec 5 composantes sont toujours meilleurs. Il est 93.3% et supérieur de 4.9% à celui utilisant le vecteur *F* en mode de reconnaissance dépendant du locuteur. En mode de reconnaissance indépendant du locuteur il est 91.5% et supérieur de 4% de celui du vecteur *F*. Mais le taux de reconnaissance utilisant le vecteur *Ftf* avec 4 composantes est inférieur très légèrement à celui utilisant le vecteur *Fetf*. Il atteint 92.9% en mode dépendant et 91.5% en mode indépendant. Cela veut dire que l'information d'énergie améliore quand même le taux de reconnaissance. Cependant,

parmi les informations supplémentaires le rôle de l'énergie est moins important que la durée et la différence de la fréquence.

- Les taux de reconnaissance des deux tests, le test1 et le test2, sont similaires. Chacun des deux tests ayant été réalisé avec des syllabes contenant des voyelles différentes, nous pouvons conclure que la reconnaissance des tons est indépendante des voyelles.
- Les taux de reconnaissance utilisant le vecteur normalisé par la fréquence fondamentale sont supérieurs à ceux utilisant le vecteur sans normalisation. Ceci est vraiment important dans le cas de la reconnaissance indépendante du locuteur.

5.2 Conclusion

La reconnaissance des tons a un rôle important dans la reconnaissance de la parole pour la langue vietnamienne, et aussi pour les langues à tons. Si les tons peuvent être reconnus correctement, alors cela permet d'éliminer l'espace de recherche des mots pour un mot inconnu, c'est-à-dire d'éliminer les combinaisons possibles entre des tons lexicaux et des base-syllabes.

Dans ce chapitre nous avons présenté un premier essai de réalisation d'un système de reconnaissance des tons vietnamiens. A partir des études sur la reconnaissance des tons du Mandarin, un vecteur caractéristique a été proposé spécialement adapté à la caractérisation des tons complexes du vietnamien. Nous montrons ainsi que la reconnaissance des tons vietnamiens est peu dépendante des voyelles. Une normalisation par une référence de base calculée avec la moyenne de l'un des deux tons monotones (ton1 ou ton2) est nécessaire. Le taux de reconnaissance est 93.3% pour la reconnaissance dépendante du locuteur et 91.5% pour la reconnaissance indépendante du locuteur.

5.3 Références

Doan T. T (1977).
Ngu am tiêng viêt (Phonétique vietnamienne)
Nha Xuat Ban Editions

Han M. S, Kim K. O (1974).
Phonetic variation of Vietnamese tones in disyllabic utterances tones
Journal of Phonetics, vol. 2, pp 223-232

Kadambe S, Boudreaux-Bartels G. F (1991)
A Comparison a Wavelet Functions for Pitch Detection of Speech Signals
Proc. IEEE ICASSP, pp 449-452.

Kadambe S, Boudreaux-Bartels G. F (1992)
Application of the Wavelet Transform for Pitch Detection of Speech Signals
IEEE Trans. Information Theory, vol. 38, no 2, pp 917-924.

Lee L-S, Tseng C-Y, Gu H-Y, Liu F-H, Chang C-H, Lin Y-H, Lee Y, Tu S-L, Hsieh S-H et Chen C- H (1993)
Golden Mandarin (I) - A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary
IEEE Trans. ASSP, vol. 1, no 2, pp 158-179.

Lee T, Ching P. C, Chan L. W., Cheng Y.H and Mak B. (1995).
Tone Recognition of isolated Cantonese syllables
IEEE Trans on Speech Audio Processing, vol. 3, No. 3, pp 204-209

Mallat S. G and Zhon S (1989).
Complete signal representation with multiscale edges
Tech. rep RRT-483-RR-219, Courant Inst. of Math. Sci..

Rabiner L. R (1989).
A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
Proceeding of th IEEE, vol. 77, no. 2, pp 257-284.

Tungthangthum A (1998)
Tone Recognition for Thai
Circuits and Systems, IEEE APCCAS 1998, Asia-Pacific Conference, p. 157-160.

Vu B. H (1999).
On the main characteristics of Vietnamese tones in their static state
Journal of Linguistic Institute of Vietnam, Vol. 6, pp 34-53

Wendt C, Petropulu A.P (1996)
Pitch determination and speech segmentation using the discrete wavelet transform.
IEEE International Symposium on Circuits and Systems, vol 2, pp 45-48.

Yang W. J, Lee J. C, Chang Y. C et Wang H. C (1988).
Hidden Markov Model for Mandarin Lexical Tone Recognition
IEEE Trans. ASSP, vol 36, no 7, pp 988-992

Young S, Kershaw D, Odell J, Ollason D, Valchev V, Woodland P (2000)
"The HTK Book".
The Cambridge University Engineering Department.

6

Reconnaissance des syllabes avec ton

La reconnaissance des syllabes avec ton des langues asiatiques, dans le mode de mots isolés, est le plus souvent réalisée par deux processus parallèles : la reconnaissance des tons et la reconnaissance des base-syllabes (syllabes indépendamment du ton). Dans le chapitre précédent, nous avons présenté une méthode de reconnaissance des tons vietnamiens, pour laquelle un modèle de Markov caché ou HMM a été utilisé avec un vecteur acoustique de 5 composantes.

Habituellement, la reconnaissance des base-syllabes est, soit réalisée avec des méthodes fondées sur les techniques de modèles HMMs, soit conduite par des méthodes de réseau neuronal ou bien par des techniques hybrides HMM/réseau neuronal. Cependant les méthodes utilisant des modèles HMMs restent les plus courantes, essentiellement dans les systèmes conçus dans les laboratoires recherches comme les systèmes HTK [Young et al, 2000] ou SPHINX [Lee et al, 1990] ou dans les systèmes commerciaux (IBM, Microsoft,...)

Pour la reconnaissance de la parole, l'unité acoustique à modéliser peut être le mot, la syllabe ou le phonème. Les unités sont modélisées indépendamment ou non du contexte. Le choix de l'unité acoustique pour la reconnaissance est dépendant des particularités de la langue et du type de système de reconnaissance.

Dans les systèmes de reconnaissance des digits anglais fondés sur un petit vocabulaire, où la base de données suffisante pour l'entraînement est facile à réaliser, l'unité acoustique choisie

est le mot, pour lequel les effets de la dépendance du contexte et de la coarticulation dans un mot sont compris dans le modèle [Rabiner et al, 1989].

Dans un système avec un grand vocabulaire, l'unité acoustique choisie est habituellement le phonème qui peut être représenté indépendamment ou non du contexte [Lee, 1990]. Lorsqu'on utilise des modèles de phonèmes, le nombre de modèles nécessaires est réduit et des phonèmes peuvent partager des paramètres. Les paramètres du modèle peuvent être alors estimés plus robustement.

La syllabe peut aussi être choisie comme l'unité acoustique pour la reconnaissance avec grand vocabulaire [Ganapathiraju et al, 2001]. Cependant, la base de données d'apprentissage est alors difficile à réaliser pour les langues occidentales comme l'anglais dont le nombre de syllabes est d'environ 30 000. Par contre, la syllabe est souvent utilisée comme unité acoustique dans le cas des langues dont le nombre de syllabes n'est pas grand comme le Mandarin qui présente environ 408 base-syllabes [Lee et al, 1993], le cantonnais 580 base-syllabes [Lee et al, 1997], le japonais et ses 114 syllabes [Nakagawa et al, 1999] et les 2400 base-syllabes du vietnamien.

Dans ce chapitre nous allons réaliser le prototype d'un système de reconnaissance de syllabes en langue vietnamienne et en mode de syllabes isolées. Pour la reconnaissance des tons la technique décrite au chapitre précédent sera utilisée. Pour la reconnaissance des base-syllabes, nous utiliserons une méthode fondée sur un modèle HMM, pour laquelle la syllabe est choisie naturellement comme l'unité acoustique à reconnaître.

Cependant, nous essayerons aussi d'utiliser des unités acoustiques qui sont des parties de la syllabe : les parties initiales et les parties finales. En effet, dans ce cas, le nombre des parties initiales et des parties finales n'est pas important et dans le cas où les données d'entraînement ne seraient pas "suffisantes", ceci pourrait permettre d'estimer plus robustement les paramètres des modèles et pourrait augmenter le taux de reconnaissance.

6.1 Corpus

Nous rappelons que le corpus utilisé pour la caractérisation des tons et la reconnaissance des tons présente seulement 13 couples de syllabes constitués de la base-syllabe et de la même

syllabe prononcée avec les tons différents, pour un total de 131 syllabes. Pour ce corpus, la reconnaissance des tons devient moins importante dans la reconnaissance des syllabes avec ton. C'est pourquoi, pour réaliser un prototype du système et évaluer la reconnaissance des syllabes nous utilisons un autre corpus qui contiendra plusieurs syllabes fondées sur la même base-syllabe mais les tons différents. Ce corpus présente 4 consonnes initiales (choisies parmi les 21 consonnes initiales possibles) qui contiennent 2 consonnes voisées /b/ et /d/ et 2 consonnes non-voisées /f/ et /c/, et 130 parties finales (choisies parmi les 155 parties finales). Les combinaisons possibles entre ces éléments et les six tons créent 1168 syllabes avec tons. Chaque syllabe est répétée quatre fois par un locuteur masculin NQC. Le nombre de syllabes du corpus est ainsi égal à 1168 * 4 = 4672 syllabes. Les données d'entraînement contiendront toutes les syllabes répétées trois fois, elles sont égales à 1168 * 3 = 3504 syllabes. Les données de test contiendront toutes les syllabes répétées une fois et différentes des données d'entraînement, c'est-à-dire 1168 syllabes. Le tableau **6.1** présente la distribution des syllabes dans le corpus, c'est à dire le nombre de base-syllabes et le nombre de tons différents pour chaque syllabe. Par exemple, la 2ème ligne montre qu'il y a 135 base-syllabes (sur un total de 413 base-syllabes) qui sont prononcées avec 2 tons pour créer 270 syllabes avec ton (sur un total de 1168 syllabes avec ton).

	nombre de base-syllabes	**nombre de tons attachés**	**nombre de syllabes avec ton**	**pourcentage de syllabes avec ton**
	92	1	92	7.9 %
	135	2	270	23.1 %
	51	3	153	13.1 %
	51	4	204	17.5 %
	55	5	275	23.5 %
	29	6	174	14.9 %
total	**413**		**1168**	**100 %**

Tableau 6.1 Nombre des syllabes avec ton du corpus

6.2 Reconnaissance des syllabes avec ton

La figure **6.1** présente le schéma d'un système de reconnaissance des syllabes avec ton qui inclut le processus de reconnaissance des base-syllabes et le processus de reconnaissance des

tons. Chaque processus va sortir les deux meilleurs candidats. Si la combinaison de la base-syllabe BS_c et le ton *i* est autorisée, alors la vraisemblance de cette syllabe sera calculée par la formule suivante :

$$P_{ST}(c,i) = P_{BS}(c).P_T(i) \tag{6.1}$$

$P_{ST}(c,i)$: vraisemblance de la syllabe avec ton

$P_{BS}(c)$: vraisemblance de la base-syllabe *c*

$P_T(i)$: vraisemblance du ton *i*

Nous rappelons qu'il existe des combinaisons impossibles de base-syllabe et de tons. Par exemple, la syllabe fermée ne peut pas se combiner avec le ton1, le ton2, le ton3 et le ton4.

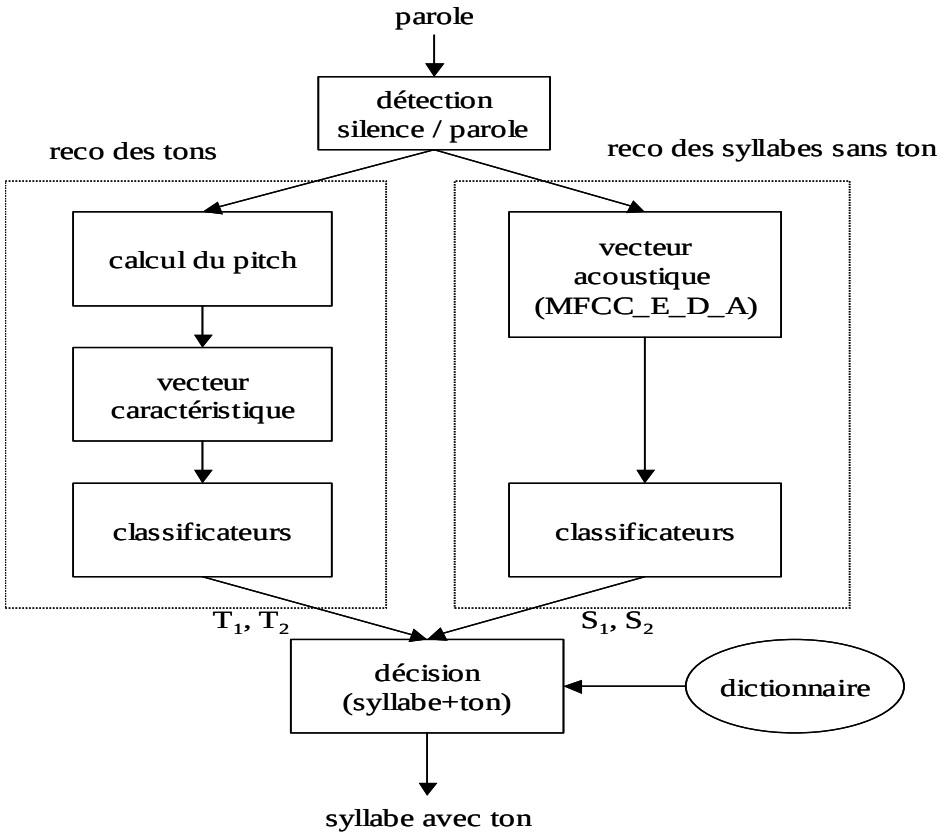


Figure 6.1 Un système complet de reconnaissance de mots (syllabes) pour le vietnamien

Les sections suivantes décrivent les parties du système complet de reconnaissance de mots de la figure **6.1**.

6.2.1 Détection parole/silence

Pour la détection parole / silence, les informations de l'énergie et du passage par zéro sont généralement utilisées [Rabiner et al, 1975][Lamel et al, 1981]. Dans notre cas, nous utilisons la méthode de [Rabiner et al, 1975] qui reste assez simple et efficace pour notre tâche de reconnaissance de mots isolés.

Soit une trame du signal {*s(1), s(2), ..., s(N)*} et soit l'énergie de la trame du signal *E(n)*:

$$E(n) = \frac{1}{N}\sum_{i=1}^{N} s^2(i) \qquad 1 \le i \le N \tag{6. 2}$$

Le taux de passage par zéro de la trame du signal s'écrit alors:

$$ZCR(n) = \sum_{i=1}^{N-1} sign\big(s(i)\big).sign\big(s(i+1)\big)$$
$$sign\big(s(i)\big) = \begin{cases} 1 & if \quad s(i) \succ 0 \\ 0 & if \quad s(i) \prec 0 \end{cases} \tag{6. 3}$$

La figure **6.2** présente l'algorithme.

La méthode utilise 3 seuils:

- ITU : seuil supérieur d'énergie ;
- ITL : seuil inférieur d'énergie ;
- IZCT : seuil de taux de passage par zéro.

Le point pour lequel l'énergie est supérieure au seuil ITU est tout d'abord recherché. Ce point est considéré comme le point dans la région du signal de parole. Pour le point au début du mot, une recherche vers l'arrière est faite. Ceci est réalisé jusqu'au point pour lequel l'énergie est inférieure au seuil ITL. Ce point est alors marqué comme le point de début temporaire N_1. Une vérification du phonème non-voisé, s'il existe, pour l'intervalle T (environ 250 ms [Rabiner et al, 1975]) sera réalisée, fondée sur le taux de passage par zéro. Si le taux de passage par zéro excède le seuil IZCT 3 fois ou plus, le point N_1 est déplacé au premier point pour lequel le seuil IZCT est franchi. Le point N_1 est alors défini comme le point au début. De la même façon le point à la fin du mot N_2 est trouvé par la recherche vers l'avant dans l'intervalle N_2+T.

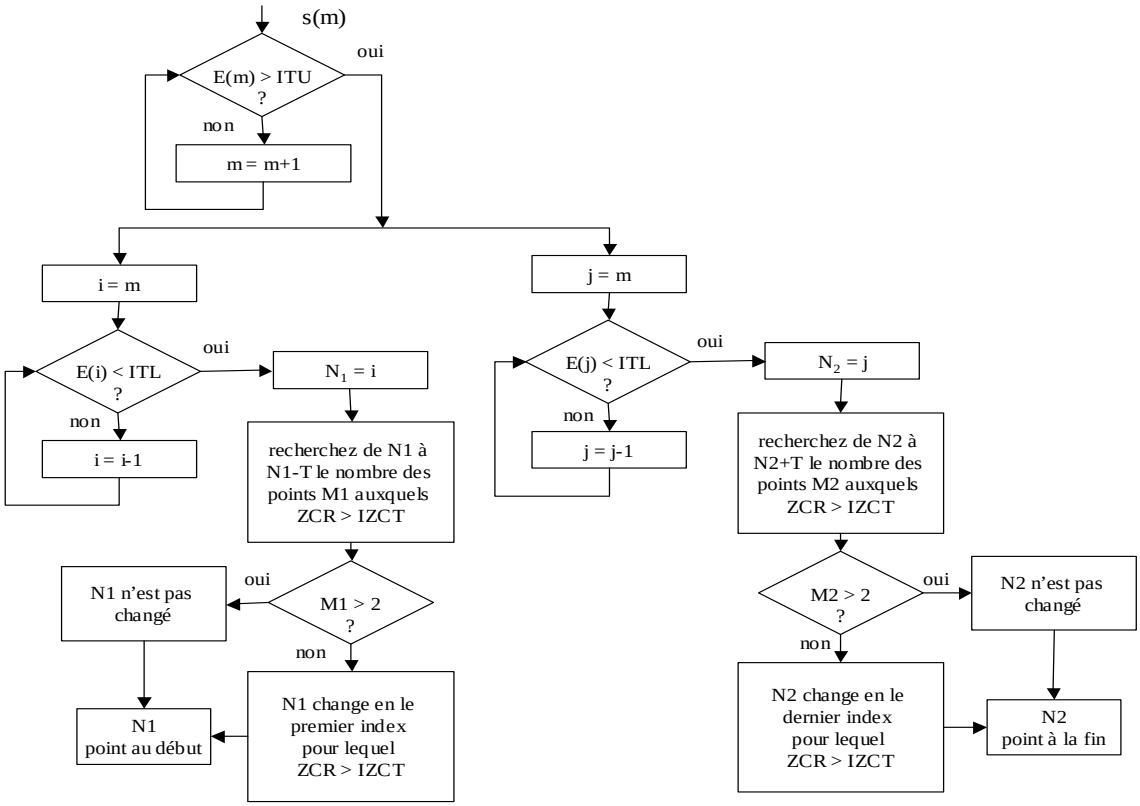


Figure 6.2 Schéma de l'algorithme de la détection parole / silence [Rabiner et al, 1975]

6.2.2 Reconnaissance des tons

Nous utilisons notre méthode présentée dans le chapitre 5 pour la reconnaissance des tons. Le vietnamien a six tons lexicaux. Mais le ton5 et le ton6 ont 2 représentations : le ton5a et le ton6a pour les syllabes ouvertes et le ton5b et le ton6b pour les syllabes fermées. Donc 8 HMMs sont utilisés donc pour 8 représentations des 6 tons. Chaque HMM contient 3 états émettant des observations. La topologie des HMMs est donnée à la figure **6.3**:

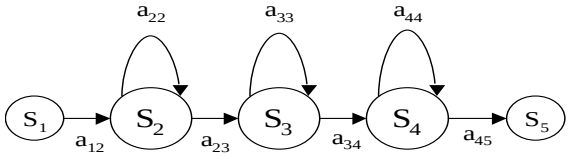


Figure 6.3 Topologie du HMM pour modéliser les tons (l'état au début et l'état à la fin n'émettent pas d'observations)

Les résultats de reconnaissance des tons pour 1 et 2 meilleurs candidats sont présentés dans le tableau **6.2**:

Taux de reconnaissance (%)	
1-meilleur	**2-meilleurs**
90.50	98.12

Tableau 6.2 Résultats de la reconnaissance des tons (nombre des tests: 1168) sur le nouveau corpus.

Nous remarquons que le taux de reconnaissance dans le cas d'un meilleur candidat est un peu moins bon que le taux de reconnaissance en mode dépendant du locuteur présenté dans le chapitre 5 qui était de 93.3%. En effet, pour valider notre méthode dans le chapitre 5, la détection de silence/parole était réalisée manuellement. De plus, quelques erreurs de calcul du pitch avaient aussi été corrigées manuellement. Dans ce chapitre tous les calculs sont réalisés automatiquement, il n'y a pas des corrections manuelles : les erreurs sont donc augmentées.

6.2.3 Reconnaissance des syllabes en utilisant la syllabe comme l'unité acoustique à reconnaître

Chaque base-syllabe est modélisée par un modèle HMM. Il y a donc 413 HMMs pour 413 base-syllabes. Les données d'entraînement sont segmentées automatique et étiquetées au niveau de la syllabe avec l'aide de la détection parole/silence. Chaque HMM est donc entraîné séparément en utilisant les données correspondantes. La topologie des HMMs est présentée à la figure **6.4**:

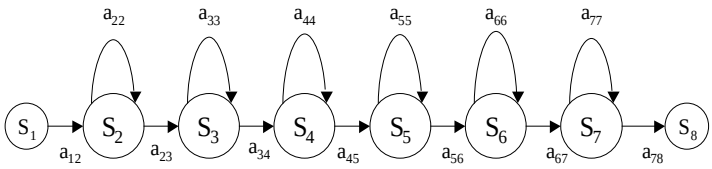


Figure 6.4 Topologie du HMM pour modéliser les base-syllabes (l'état au début et l'état à la fin n'émettent pas d'observations)

Le résultat de la reconnaissance des base-syllabes est présenté dans le tableau **6.3** pour différent nombre de gaussiennes pour chaque état du HMM:

	Taux de reconnaissance (%)	
nombre des gaussiennes	**1-meilleur**	**2-meilleurs**
1	79.54	88.10
2	72.17	80.14
4	66.01	74.04

Tableau 6.3 Taux de reconnaissance des base-syllabes en utilisant des modèles de syllabes (nombre des tests: 1168)

Le taux de reconnaissance diminue lorsque le nombre des gaussiennes augmentent, c'est-à-dire que le nombre de paramètres à estimer augmente. Ceci peut être expliqué par le fait que dans ce cas les paramètres ne sont pas estimés correctement car les données d'entraînement sont limitées.

Le résultat du couplage de la reconnaissance des tons et de la reconnaissance des base-syllabes pour créer 1168 syllabes avec ton est donné dans le tableau **6.4** :

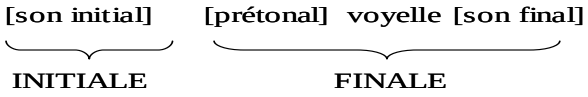
Nombre des gaussiennes	Taux de reconnaissance (%)
1	*74.06*
2	*67.38*
4	*60.96*

Tableau 6.4 Taux de reconnaissance de mots (ou syllabes avec ton) (nombre des tests: 1168)

Au vu de ces résultats, nous pouvons conclure que la modélisation de base-syllabes n'est pas adéquate lorsque les données sont limitées puisque plus de 400 modèles doivent alors être entraînés. La section suivante propose donc de modéliser un autre type d'unité acoustique.

6.2.4 Reconnaissance des syllabes en utilisant la structure INITIALE/FINALE

La base-syllabe du vietnamien peut être divisée en deux parties: la partie initiale et la partie finale, comme nous l'avons présenté au chapitre 2 de notre mémoire.



La partie initiale peut être un phonème et la partie finale peut comprendre de 1 à 3 phonèmes. Il est alors possible d'utiliser les parties initiale et finale comme les unités acoustiques à reconnaître.

Pour la reconnaissance des syllabes isolées, il n'y a pas d'effets de coarticulation entre les syllabes. De plus, dans une syllabe, si la partie initiale est très dépendante de la partie finale, la partie finale est, quant à elle, faiblement dépendante de la partie initiale. La partie initiale peut donc être modélisée par un modèle dépendant du contexte droit et la partie finale peut être modélisée par un modèle indépendant du contexte.

On peut considérer que la partie initiale dépend seulement du phonème placé à sa droite, c'est-à-dire qu'elle est dépendante du prétonal s'il existe, ou de la voyelle, mais pas de toute la partie finale. Par exemple, la syllabe {ban} peut être modélisée par un modèle dépendant du contexte droit de la partie initiale {b+a} et par un modèle indépendant du contexte pour la partie finale {an} au lieu de l'utilisation du modèle {b+an} et du modèle {an}. Ceci va réduire le nombre de modèles du système. Les paramètres de chaque modèle pourront alors être estimés plus robustement, sur une quantité de données plus importante.

Les 413 base-syllabes sont modélisées par 67 modèles initiaux dépendant du contexte droit et 130 modèles finaux indépendants du contexte. Nous utilisons des HMMs avec 3 états émettant des observations pour la partie initiale et des HMMs à 5 états émettant des observations pour la partie finale.

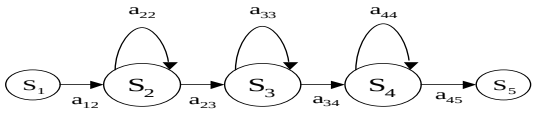


Figure 6.5 Topologie du HMM pour le modèle initial

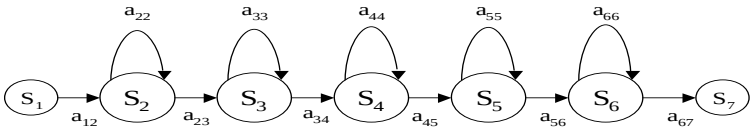


Figure 6.6 Topologie du HMM pour le modèle final

Pour la reconnaissance en utilisant la structure INITIALE/FINALE, les données d'entraînement sont segmentées et étiquetées seulement au niveau de la syllabe par l'utilisation de la détection parole/silence. Cela signifie que nous savons que le segment du signal considéré correspond à une syllabe mais nous ne savons quelle portion de signal correspond à la partie initiale et quelle portion correspond à la partie finale. L'entraînement du HMM ne peut donc pas être réalisé séparément pour les parties initiales et finales. Nous avons donc besoin d'utiliser l'approche consistant à estimer les paramètres des modèles qui est appelée "embedded training" en utilisant l'algorithme Baum-Welch. D'abord la concaténation des modèles des parties initiales et des modèles des parties finales pour créer les modèle des syllabes est réalisée. Puis nous considérons le modèle combiné comme un "gros" modèle de syllabe, les probabilités concernant chaque état du modèle sont calculées en utilisant l'algorithme avant-arrière. Ces probabilités seront mémorisées pour chaque état. Tous les "gros" modèles de syllabe sont traités. A la fin, les modèles de la partie initiale et de la partie finale seront estimés à partir des probabilités mémorisées de chaque état du modèle.

Le taux de reconnaissance des base-syllabes en utilisant les modèles initiaux et les modèles finaux est présenté dans le tableau **6.5**

	Taux de reconnaissance (%)	
nombre des gaussiennes	1-meilleur	2-meilleurs
1	82.19	89.90
2	85.70	92.21
4	85.70	91.78

Tableau 6.5 Taux de reconnaissance des base-syllabes en utilisant la structure INITIALE/FINALE (nombre des tests: 1168)

Le taux de reconnaissance est plus grand que celui utilisant les modèles de syllabe, et il augmente avec l'augmentation du nombre de gaussiennes. Ceci veut dire que les modèles sont appris avec un peu plus de données d'entraînement, les paramètres sont donc estimés plus robustement.

Les résultats du couplage de la reconnaissance des tons et de la reconnaissance des base-syllabes pour créer 1168 syllabes avec ton sont donnés dans le tableau **6.6**

Nombre des gaussiennes	Taux de reconnaissance (%)
1	*77.83*
2	*80.74*
4	*80.31*

Tableau 6.6 Taux de reconnaissance de mots ou syllabes avec ton (nombre des tests: 1168)

6.3 Discussions et Conclusions

Le taux de reconnaissance de la base-syllabe en modélisant les parties initiales et finales comme unités acoustiques à reconnaître est supérieur d'environ 6% pour le cas de 1 meilleur candidat que celui obtenu en utilisant la syllabe comme l'unité acoustique à reconnaître. Ceci peut être expliqué par le fait que, dans le cas de données d'entraînement insuffisantes (chaque syllabe avec ton est répétée seulement 3 fois pour l'entraînement), l'utilisation du modèle initial dépendant du contexte et du modèle final indépendant du contexte permet de réduire le nombre des modèles nécessaires. Le modèle reçoit alors un peu plus de données d'entraînement et les paramètres du modèle sont estimés plus robustement.

Le taux de reconnaissance atteint 85,7 % pour la reconnaissance des base-syllabes et 80,7% pour la reconnaissance des syllabes avec ton. Ceci est comparable aux résultats des systèmes du Mandarin ou du cantonnais pour la reconnaissance des syllabes isolées [Liu et al, 1993][Lee, 1997].

Un système de reconnaissance du cantonnais utilisant une méthode fondée sur des techniques de réseau neuronal en mode dépendant du locuteur a été réalisé par [Lee, 1997]. Le vocabulaire contient 200 syllabes avec ton comprenant 166 syllabes sans ton avec 3 locuteurs, chaque syllabe est répétée 12 fois par chaque locuteur. La moitié des données est utilisée pour l'entraînement et le reste est utilisé pour les tests. Le taux de reconnaissance des syllabes avec ton obtenu par Lee est alors de 81,8%.

Un système de reconnaissance de syllabes sans ton du Mandarin utilisant une méthode fondée sur des modèles HMMs en mode dépendant du locuteur est réalisé par [Liu et al, 1993]. Le vocabulaire contient 408 syllabes avec seulement le ton1 du Mandarin. Chaque syllabe est prononcée 6 fois par chaque locuteur. Il y a 2 locuteurs. Les données d'entraînement pour

chaque locuteur se composent de cinq exemplaires pour chacune des 408 syllabes, l'exemplaire restant est utilisé pour la phase de test. Le taux de reconnaissance des syllabes sans ton atteint 90,6%. Néanmoins, dans ce cas, les données d'entraînement ont dû être segmentées en une partie initiale et une partie finale pour chaque syllabe, ce qui est difficile à réaliser automatiquement, spécialement dans la syllabe dont la partie initiale est voisée. C'est pourquoi, dans l'étude de [Liu et al, 1993] les segmentations INITIALE/FINALE ont été réalisées manuellement. Si la segmentation INITIALE/FINALE n'était pas utilisée, alors le taux de reconnaissance des syllabes sans ton ne serait que 82.1%.

Notre taux de reconnaissance des syllabes avec ton 80.7% n'est pas totalement satisfaisant pour la reconnaissance des mots isolés en mode dépendant du locuteur. Cependant, ce taux est acceptable si on prend en compte le fait que les données d'entraînement sont limitées (chaque syllabe est répétée seulement 3 fois pour l'entraînement).

6.4 Références

Doan T. T (1977).
Ngu am tiêng viêt (Phonétique vietnamienne)
Nha Xuat Ban Editions

Ganapathiraju A, Hamaker J, Picone J, Ordowski M, Doddington G.R (2001)
Syllable-Based Large Vocabulary Continuous Speech Recognition
IEEE Trans on Speech and Audio Processing, Vol 9, No 4,pp 358-366

Le K.F (1990)
An Overview of the Sphinx Speech Recognition System
IEEE Trans on ASSP, Vol. 38, No. 1, pp 35-45

Lee K.F (1990)
Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition
IEEE Trans on ASSP, Vol 38, No 4, pp 599-609

Lee T, Ching P.C (1999)
Cantonese Syllable Recognition Using Neural Networks
IEEE Trans on Speech and Audio Processing, Vol 7, No 4, pp 466-472

Liu F-H, Lee Y, Lee L-S (1993)
A Direct-Concatenation Approach to Train Hidden Markov Models to Recognize the Highly Confusing Mandarin Syllables with Very Limited Training Data
IEEE Trans on Speech and Audio Processing, Vol 1, No 1, pp 113-119

Nakagawa S, Hanai K, Yamamoto K, Minematsu N (1999)
Comparison of syllable-based HMMs and Triphone-based HMMs in Japanese Speech Recognition
ASRU 1999

Rabiner L.R, Sambur M.R (1975)
An algorithm for determining the endpoints of isolated utterances
Bell System Technical Journal, Vol 54, pp 297-315

Rabiner L. R (1989).
A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
Proceeding of th IEEE, vol. 77, no. 2, pp 257-284.

Rabiner L.R, Wilpon J.G, Soong F.K (1989)
High Performance Connected Digit Recognition Using Hidden Markov Models
IEEE Trans on ASSP, Vol 37, No 8, pp 1214-1225

Schwartz R, Austin S (1991)
"A Comparison of Several Approximate Algorithms For Finding Multiple (N-Best) Sentence Hypotheses"
Proc. IEEE ICASSP 1991, pp 701-704

Young S, Kershaw D, Odell J, Ollason D, Valchev V, Woodland P (2000)
"The HTK Book"
The Cambridge University Engineering Department.

7

Conclusions et perspectives

7.1 Conclusions

La reconnaissance de la parole est une partie importante dans un système de dialogue homme-machine ainsi que dans un système de traduction automatique multilingue de la parole. Nos objectifs étaient d'étudier les caractéristiques linguistiques et acoustiques du vietnamien pour la reconnaissance de la parole et de réaliser un système de reconnaissance automatique de cette langue en mode de syllabe isolée. Ces études sont nécessaires dans un contexte de recherche de plus en plus multilingue où le vietnamien n'a été que très peu abordé jusqu'à maintenant.

Dans le deuxième chapitre de notre mémoire, nous avons présenté les caractéristiques de la phonologie et de l'acoustique du vietnamien, d'une manière générale mais en essayant de faire ressortir celles qui pourraient nous servir pour la reconnaissance. Le vietnamien est une langue monosyllabique. Une syllabe peut être analysée en deux parties : la partie initiale et la partie finale. Il y a 22 parties initiales et 155 parties finales. Une liste complète de ces parties a été présentée, ce qui nous a permis de faire la transcription phonologique des syllabes pour la reconnaissance en utilisant la structure initiale/finale.

Le vietnamien est une langue tonale. Les tons du vietnamien ont été beaucoup étudiés par les linguistes essentiellement, mais d'une manière uniquement fondée sur la perception auditive. Par contre, très peu de travaux ont étudié les tons au niveau des expérimentations pratiques.

C'est pourquoi, il n'existe pas de corpus du vietnamien en libre accès. Nous avons donc réalisé un corpus avec 15 locuteurs. Ce corpus contient un ensemble de syllabes présentant les 16 voyelles et les 21 consonnes initiales. Chaque voyelle est combinée au minimum une fois avec chacun des 6 tons. Le corpus a une durée d'environ 105 minutes.

Grâce à ce corpus, nous avons caractérisé les six tons de la langue vietnamienne, comme décrit dans le chapitre 4. Les gabarits et les durées des tons sont présentés. Nous avons alors comparé nos résultats, nos remarques extraites à partir de nos données, aux différentes études précédentes disponibles et nous avons constaté qu'ils étaient concordants.

Pour les langues tonales, la reconnaissance de ton est une tâche importante. L'utilisation de paramètres concernant la variation du pitch va de soi pour la reconnaissance des tons. Un vecteur avec 2 composantes du pitch (amplitude et pente) a été appliqué essentiellement à la reconnaissance des 4 tons lexicaux du mandarin. Mais ce vecteur n'est pas performant pour les 6 tons lexicaux vietnamiens avec 8 représentations. Pour augmenter la performance, nous avons donc essayé d'ajouter alternativement les informations supplémentaires de l'énergie, et de la durée et de la différence entre le point terminal et le point au milieu du ton. Nous avons trouvé qu'un vecteur avec 5 composantes (l'amplitude locale, la pente locale, la durée, l'énergie et la pente globale), donne le meilleur résultat. Le taux de reconnaissance atteint est de 93.3% en mode dépendant du locuteur et 91.5% en mode indépendant du locuteur. Nous avons également montré qu'une normalisation de la valeur du pitch est nécessaire dans le mode de reconnaissance indépendante du locuteur. Notre système de reconnaissance des tons du vietnamien semble par ailleurs peu dépendant des voyelles.

Le problème à résoudre pour la reconnaissance de base-syllabes (syllabes indépendamment du ton) du vietnamien n'est pas différent des autres langues comme l'anglais, le français ou le mandarin. C'est-à-dire que les techniques existantes peuvent être appliquées au vietnamien. La technique du HMM est beaucoup développée et utilisée largement dans la reconnaissance de la parole. Nous avons donc appliqué cette technique à la reconnaissance du vietnamien avec un dictionnaire de 1168 syllabes. Le vietnamien est monosyllabique et le nombre de monosyllabes n'est pas grand, environ 2400 base-syllabes. L'unité acoustique à reconnaître choisie est naturellement la syllabe. Néanmoins, dans le cas de données insuffisantes, l'utilisation du modèle de syllabe n'est pas performante. Une utilisation des modèles de la partie initiale et de la partie finale issus de notre connaissance de la langue vietnamienne, pour

réduire le nombre des modèles, est donc plus efficace. Dans nos expériences, le taux de reconnaissance de base-syllabes est 79.5% en utilisant comme unité acoustique la syllabe. Il devient 85.7% en utilisant les modèles de la partie initiale et de la partie finale.

Nous avons couplé la partie de reconnaissance des tons et la partie de reconnaissance de syllabe sans ton pour faire un prototype du système de reconnaissance du vietnamien en mode de mot isolé. Le taux de reconnaissance atteint est de 80.74%.

7.2 Perspectives

Dans le future, pour atteindre un système complet de reconnaissance de syllabes en mode isolé et indépendant du locuteur, il sera nécessaire de constituer un corpus comprenant toutes les syllabes, avec plusieurs locuteurs et il faudra aussi envisager l'application des techniques d'adaptation au locuteur.

Si l'on veut passer à la reconnaissance automatique de la parole continue, il y a plusieurs problèmes à résoudre.

Les consonnes finales du vietnamien sont les consonnes implosives. La relation entre les syllabes est donc assez discrète. La frontière des syllabes dans la parole n'est donc pas trop difficile à détecter. La reconnaissance de la syllabe isolée décrite dans cette thèse pourrait être utilisée en reconnaissance de parole continue comme la partie de décodage acoustique. Néanmoins nous avons vu que le contour d'un ton peut être dépendant du ton précédent et du ton suivant. Une étude de la dépendance des tons en contexte est donc nécessaire.

En reconnaissance de parole continue, le décodage acoustique ne donne pas des résultats fiables à 100%. Nous avons donc besoin d'un modèle de langage. Le modèle de langage peut contenir une grammaire ou des modèles de langage stochastiques comme les N-grammes. Pour cela, une étude de la syntaxe, ainsi que la construction d'une base de données de textes du vietnamien sont nécessaires. Pour les données de textes, on peut envisager de les collecter à partir de plusieurs sources disponibles: les journaux, les livres ou bien les newsgroups et les pages Web sur Internet.

La reconnaissance automatique de la parole est normalement performante lorsque les conditions environnementales et sonores de l'utilisation du moteur de reconnaissance sont proches des conditions environnementales et sonores constatées lors de la phase d'apprentissage. Ceci est malheureusement rarement obtenu car le signal de parole est très souvent influencé par le nombre de sources de bruit venant de l'environnement et leur nature, le type de microphone, l'écho de la salle, ou bien la distorsion de la transmission (en cas d'utilisation du téléphone par exemple).

Le problème *de robustesse* de la reconnaissance est encore un problème fortement étudié de nos jours, car pas encore complètement résolu. Plusieurs techniques peuvent être proposées : l'adaptation du moteur de reconnaissance aux nouvelles conditions environnementales, l'utilisation d'informations supplémentaires comme des informations visuelles sur la géométrie des lèvres ou bien des techniques de traitement du signal appelées généralement *rehaussement* du signal de parole.

Le rehaussement du signal de parole par séparation aveugle de sources sonores est l'une des méthodes. Nous avons étudié cette technique au début de notre travail de thèse (travail décrit dans l'annexe) mais nous n'avons pas eu l'occasion de l'appliquer à notre système de reconnaissance du vietnamien. A court terme, nous envisageons d'utiliser cette technique (et éventuellement d'autres techniques de traitement du signal complémentaires), en amont, comme un module de pré-traitement du signal de parole, dans notre système de reconnaissance du vietnamien.

Références bibliographiques globales

Andreev N.D et Gordina M.V (1957)
Н. Д. Андреев et М. В. Гордина
Système des tons vietnamiens (en russe)
по эсперименталъым данным, Весник, No 8

Atal B et Hanauer S (1971)
Speech analysis and synthesis by linear prediction of the speech wave
Journal of the Acoustic Soc. Am, vol 50, pp 637-655.

Bagshaw P.C, Hiller S.M, Jack M.A (1993)
*Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. (*http://www.cstr.ed.ac.uk/~pcb/fda_eval.tar.gz*)*
Proceedings of EuroSpeech'93, Berlin.

Barnett J, Bamberg P, Held M, Huerta J, Manganaro L, Weiss A (1995)
Comparative Performance in Large-Vocabulary Isolated-Word Recognition in Five European Languages
EuroSpeech'95, pp 189-192

Baum L. E. (1972).
An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes
Inequalities, vol. 3, pp 1-8,

Bellik Y (1997)
Multimodal Text Editor Interface Including Speech for the Blind
Speech Communication, vol 23, No. 4, Decmber

Boite R., Bourlard H., Dutoit T., Hancq J., Leich H. (2000)
Traitement de la parole
Presse Polytechniques et Universitaires Romandes.

Boitet C, Caelen J, Fafiotte, G, Keller, E, Lafourcade, M, Wehrli, E (1998)
Integrating French within C-STAR II
Grenoble,Report and demos of the CLIPS++ group

Calliope (1989)
La Parole et son Traitement Automatique
Editions Masson - Paris.

Chang E, Zhou J, Di S, Huang C, Lee K-F (2000)
Large vocabulary Mandarin speech recognition with different approaches in modeling tones
6th International Conference of Spoken Language Processing, Beijing

Chen C. J, Li H, Shen L, Fu G (2001)
Recognize tone languages using pitch information on the main vowel of each syllable
In Proc of ICASSP.

Chen C.J, Gopinath R. A, Monkowski M. D, Picheny M. A et Shen K (1997)
New methods in continuous Mandarin Speech Recognition
Eurpspeech'97, pp 1543-1546

Davis S.B., Mermelstein P. (1980)
Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,
IEEE Trans. on ASSP, vol. 28, no. 4, pp 357-366,.

Doan T.T (1977).
Ngữ âm tiếng Việt (Phonétique vietnamienne)
Nhà xuất Đại học và Trung học chuyên nghiệp

Forney G. D. (1973)
Viterbi algorithm
Proccedings of the IEEE, vol. 61, pp 268-278.

Gales M.J.F (2001)
Adaptive Training for Robust ASR
Automatic Speech Recognition and Understanding Workshop, ASRU2001

Ganapathiraju A, Hamaker J, Picone J, Ordowski M, Doddington G.R (2001)
Syllable-Based Large Vocabulary Continuous Speech Recognition
IEEE Trans on Speech and Audio Processing, Vol 9, No 4,pp 358-366

Gold. B et Rabiner. L (1969)
Parallel processing techniques for estimating pitch periods of speech in the time domain.
Acoustical Society of America, 46(2), pp442-448.

Grimes B. F
Ethnologue: Languages of the World
Dallas, Texas: Summer Institute of Linguistics

Han M. S, Kim K. O (1974).
Phonetic variation of Vietnamese tones in disyllabic utterances tones
Journal of Phonetics, vol. 2, pp 223-232

Haudricourt A.G (1953)
La place du vietnamien dans les langues austroasiatiques
Bulletin de la Société de Linguistique de Paris, 1953, 49, 1

Haudricourt A.G (1954)
De l'origine des tons en vietnamiens
Journal Asiatique, 1954, 242, 1

Hermansky H. (1990)
Perceptual Linear Predictive (PLP) analysis of speech
Journal of the Acoustic Soc. Am, vol. 87, no. 4, pp 1738-1752.

Hess. W.H (1983)
Pitch Determination of Speech Signal: Algorithms and Devices.
Springer-Verlag, Heidelberg, Germany.

Hoang T, Hoang M (1975)
Remarques sur la structure phonologique du vietnamien
Etudes des vietnamiens, No 40, Hanoi

Hwang, M.Y., X. Huang, and F. Alleva (1993)
Predicting Unseen Triphones with Senones
ICASSP' 93, pp 311-314

Igounet S. (1998)
Eléments pour un système de reconnaissance automatique de la parole continue du français
Thèse Informatique, Université d'Avignon et des Pays de Vaucluse .

Itakura F et Saito S (1968)
Analysis synthesis telephony based upon the maximum likelihood method
In Kohasi Y, editor, 6th International Congress on Acoustics, Tokyo, pages C-5-5.

Jelinek F (1976)
Continuous Speech Recognition by Statistical Methods
Proc of the IEEE, vol 64, No 4, pp 532-557

Juang B.H (1998)
The Past, Present, and Future of Speech Processing
IEEE Signal Processing Magazine, May 1998, pp 24-48

Kadambe S, Boudreaux-Bartels G. F (1991)
A Comparison a Wavelet Functions for Pitch Detection of Speech Signals.
Proc. IEEE ICASSP, pp 449-452.

Kadambe S, Boudreaux-Bartels G. F (1992)
Application of the Wavelet Transform for Pitch Detection of Speech Signals.
IEEE Trans. Information Theory, vol. 38, no 2, pp 917-924.

Karoonboonyanan T (1999)
Standardization and Implementations of Thai Language
presented at the Seminar on Enhancement of the International Standardization Activities in Asia Pacific Region **(AHTS-1)** held on at CICC, Japan, in March 1999.

Klatt D.H. (1977)
Review of the ARPA Speech Understanding Project
JASA, Vol. 62, n°6, pp.1345-1366

Lamel L. F, Adda G, Adda-Decker M (1996)
Les lexiques de prononciation dans les systèmes de reconnaissance de la parole
Séminaire GDR-PRC, Lexique et communication parlée, Toulouse, pp 1-10

Le K.F
An Overview of the Sphinx Speech Recognition System
IEEE Trans on ASSP, Vol. 38, No. 1, pp 35-45

Le V.L (1948)
Le parler vietnamien
Paris, 1948

Lee K.F (1990)
Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition
IEEE Trans on ASSP, Vol 38, No 4, pp 599-609

Lee L-S (1997)
Voice Dictation of Mandarin Chinese
IEEE Signal Processing Magazine, July 1997

Lee L-S, Tseng C-Y, Gu H-Y, Liu F-H, Chang C-H, Lin Y-H, Lee Y, Tu S-L, Hsieh S-H et Chen C-H (1993)
Golden Mandarin (I) - A Real Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary
IEEE Trans. on Speech and Audio Processing, vol. 1, no. 2, pp 158-179.

Lee L-S, Tseng C-Y, Gu H-Y, Liu F-H, Chang C-H, Lin Y-H, Lee Y, Tu S-L, Hsieh S-H et Chen C- H (1993)
Golden Mandarin (I) - A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary.
IEEE Trans. ASSP, vol. 1, no 2, pp 158-179.

Lee L-S. (1997)
Voice Dictation of Mandarin Chinese
IEEE Signal Processing Magazine, pp 63-101.

Lee T, Ching P. C, Chan L. W., Cheng Y.H and Mak B. (1995).
Tone Recognition of isolated Cantonese syllables
IEEE Trans on Speech Audio Processing, vol. 3, No. 3, pp 204-209

Lee T, Ching P.C (1999)
Cantonese Syllable Recognition Using Neural Networks
IEEE Trans on Speech and Audio Processing, Vol 7, No 4, pp 466-472

Liu F-H, Lee Y, Lee L-S (1993)
A Direct-Concatenation Approach to Train Hidden Markov Models to Recognize the Highly Confusing Mandarin Syllables with Very Limited Training Data
IEEE Trans on Speech and Audio Processing, Vol 1, No 1, pp 113-119

Lyu R-Y, Chien L-F Hwang S-H, H H-Y, Yang R-C, Bai B-R, Weng J-C, Yang Y-J, Lin S-W, Chen K-J, Tseng C-Y, Lee L-S (1995)
Golden Mandarin (III) - A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary
In Proc of ICASSP, pp 57-60

Makhoul J. (1975)
Linear Prediction: A Tutorial Review
Proceedings of the IEEE, vol. 63, no. 4, pp 561-580.

Mallat S. G. et S. Zhong (1989)
Complete signal representation with multi scale edges.
Tech. rep RRT-483-RR-219, Courant Inst. of Math. Sci., Dec. 1989.

Medan Y, Yair. E et Chazan D (1991)
Super resolution pitch determination of speech signals.
IEEE Trans. Signal Processing, ASSP vol 39(1), pp 40-48.

Nakagawa S, Hanai K, Yamamoto K, Minematsu N (1999)
Comparison of syllable-based HMMs and Triphone-based HMMs in Japanese Speech Recognition
ASRU 1999

Newell A., Barnett J., Forgie J.W., Green C.C., Klatt D.H.,. Licklider J.C.R, Munson J., Reddy D.R. & Woods W.A. (1973)
Speech Understanding Systems: Final Report of a Study Group
North-Holland/American Elseveir, Amsterdam.

Nguyen Thi H. L (1993)
Séparation aveugle de Sources à large bande dans un mélange convolutif: Application au rehaussement de la parole
Thèse doctorat de l'INP Grenoble

Noll A.M (1967)
Cepstrum pitch determination.
Journal of the Acoustical Society of America, vol 41(2), pp 239-309.

Noll A.M (1970)
Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate.
Symposium on Computer Processing in Communication, vol 19, pp 779-797. Poly-technique Institute of Brooklyn Microwave Research Institute, New York.

Paescler A et Ney H (1989)
Continuous speech recognition using a stochastic language model
In Proc of ICASSP, pp 699-702.

Phillips M.S (1985)
A feature-based time domain pitch tracker.
Journal of the Acoustical Society of America, vol 77: S9-S10(A)

Picone J. W (1993)
Signal Modeling Techniques in Speech Recognition
Proc of the IEEE, vol. 81, No. 9, pp 1215-1247

Potisuk S, Harper M.P, Gandour J (1997)
Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method
IEEE Trans on Speech and Audio Processing, vol 7, No 1, pp 95-102

Rabiner L,.R, Cheng M.J, Rosenberg A.E and McGonegal C.A (1976)
A Comparative performance study of several pitch detection algorithms.
IEEE Trans. Audio, Signal and Speech Processing, vol 24, pp 399-417.

Rabiner L. R (1989).
A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
Proceeding of th IEEE, vol. 77, no. 2, pp 257-284.

Rabiner L. R, Juang B H (1993)
Fundamentals of Speech Recognition
published by Prentice Hall.

Rabiner L.R, Wilpon J.G, Soong F.K (1989)
High Performance Connected Digit Recognition Using Hidden Markov Models
IEEE Trans on ASSP, Vol 37, No 8, pp 1214-1225

Rabinur L.R, Sambur M.R (1975)
An algorithm for determining the endpoints of isolated utterances
Bell System Technical Journal, Vol 54, pp 297-315

Schafer R.W (1994)
in Voice Communication between Humans and Machines
National Academy Press, Washington D.C 1994

Schroeder M.R (1968)
Period histogram and product spectrum: New methods for fundamental frequency measurement.
Journal of the Acoustical Society of America, vol 43(4), pp 829-834.

Schwartz R, Austin S (1991)
A Comparison of Several Approximate Algorithms For Finding Multiple (N-Best) Sentence Hypotheses
Proc. IEEE ICASSP 1991, pp 701-704

Secrest B.G et Doddington (1983)
An integrated pitch tracking algorithm for speech systems.
Proc. IEEE ICASSP-83, pp 1352-1355, Boston.

Shen J-L (1998)
Continuous Mandarin Speech Recognition for Chinese language with large vocabulary based on segmental probability model
IEE Proc-Vis. Image Signal Process, vol. 145, No. 5, pp 309-315.

Shen J-L, Wang H-M, Bai B-R et Lee L-S (1994)
An Initial Study on A Segmental Probability Model Approach to Large-Vocabulary Continuous Mandarin Speech Recognition
In Proc of ICASSP, pp 133-136

Spalanzani A (1999)
Algorithmes évolutionnaires pour l`étude de la robustesse des systèmes de reconnaissance automatique de la parole
Thèse, Université Joseph Fourier

Thompson L.C (1965)
A Vietnamese Grammar
University of Washington Press, Seattle, 1965

Tungthangthum A (1998)
Tone Recognition for Thai
Circuits and Systems, IEEE APCCAS 1998, Asia-Pacific Conference, p. 157-160.

Viterbi A.J. (1967)
Error bounds for convolutional codes and an asymptotically optimal decoding algorithm
IEEE Trans. Informat. Theory, vol. IT-13, pp 260-269.

Vu B. H (1999).
On the main characteristics of Vietnamese tones in their static state
Journal of Linguistic Institute of Vietnam, Vol. 6, pp 34-53

Vu K. B (1984)
Untersuchungen zu den wesentlichen akustischen parametern der vietnamesischen silben (Grundfrequenz - Intensitätsverlauf und Dauer)
Études aux paramètres acoustiques essentiels des syllabes vietnamiennes (fréquence de base - cours d'intensité et durée)
Phill. Diss, HU, Berlin, 1984

Wang H-M, Ho T-H, Yang R-C, Shen J-L, Bai B-R, Hong J-C, Chen W-P, Yu T-L, Lee L-S (1997)
Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data
IEEE Trans on Speech and Audio Processing, Vol. 5, No. 2, pp195-200.

Wang H-M, Shen J-L, Yang Y-J, Tseng C-Y et Lee L-S (1995)
Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data
In Proc of LCASSP, pp 61-64

Wendt C, Petropulu A.P (1996)
Pitch determination and speech segmentation using the discrete wavelet transform.
IEEE International Symposium on Circuits and Systems, vol 2, pp 45-48.

Wendt C, Petropulu A.P (1996)

Pitch determination and speech segmentation using the discrete wavelet transform.
IEEE International Symposium on Circuits and Systems, vol 2, pp 45-48.

Xu Y (1997)
Contextual tonal variations in Mandarin
Journal of Phonetics, vol 25, pp 61-83

Yang W. J, Lee J. C, Chang Y. C et Wang H. C (1988).
Hidden Markov Model for Mandarin Lexical Tone Recognition
IEEE Trans. ASSP, vol 36, no 7, pp 988-992

Yang W. J, Lee J. C, Chang Y. C et Wang H. C (1988).
Hidden Markov Model for Mandarin Lexical Tone Recognition
IEEE Trans. ASSP, vol 36, no 7, pp 988-992

Young S, Kershaw D, Odell J, Ollason D, Valchev V, Woodland P (2000)
The HTK Book.
The Cambridge University Engineering Department.

Young S.J, Woodland P.C (1994)
State Clustering in HMM-based Continuous Speech Recognition
Computer Speech and Language, Vol 8, No 4, pp 369-384

Young Steven., (1996)
A Review of Large-vocabulary Continuous-speech Recognition
IEEE Signal Processing Magazine, pp 45-57.

Web2000:
http://www.seasite.niu.edu/vietnamese/VNMainPage/vietsite/vietsite.htm

Web ThaiARC:
http://thaiarc.tu.ac.th/host/thaiarc/thai

Web 2002:
http://www.public.asu.edu/~ickpl/

Annexes

Annexe A

Gabarits des tons

Dans cette annexe, nous présentons les gabarits des six tons que nous avons caractérisés dans le chapitre 4 pour chaque sujet.

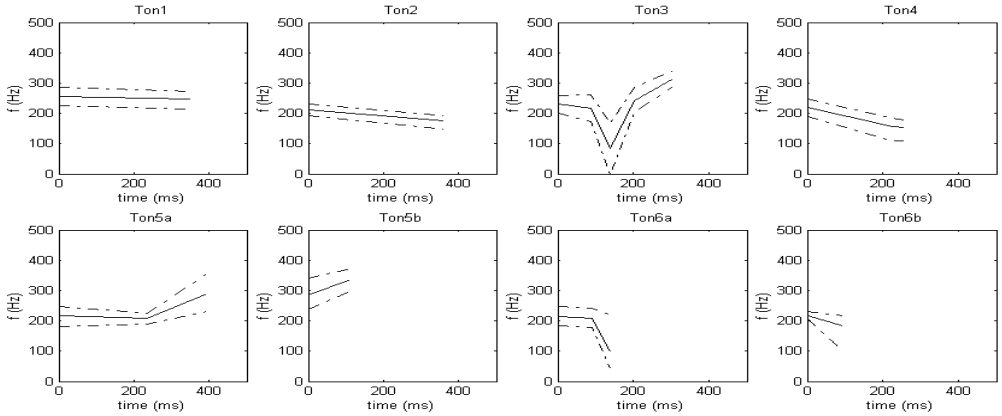


Figure A.1 Gabarits des six tons du sujet féminin PNY du Nord

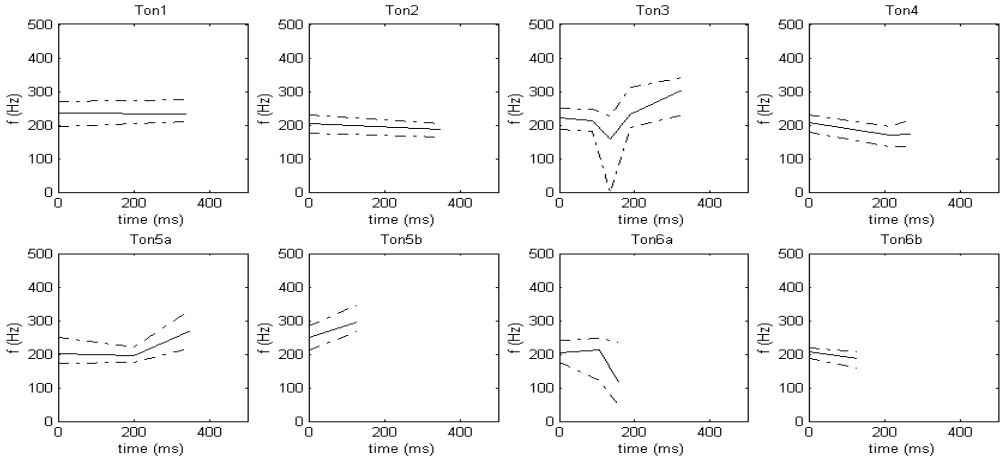


Figure A.2 Gabarits des six tons du sujet féminin VTT du Nord

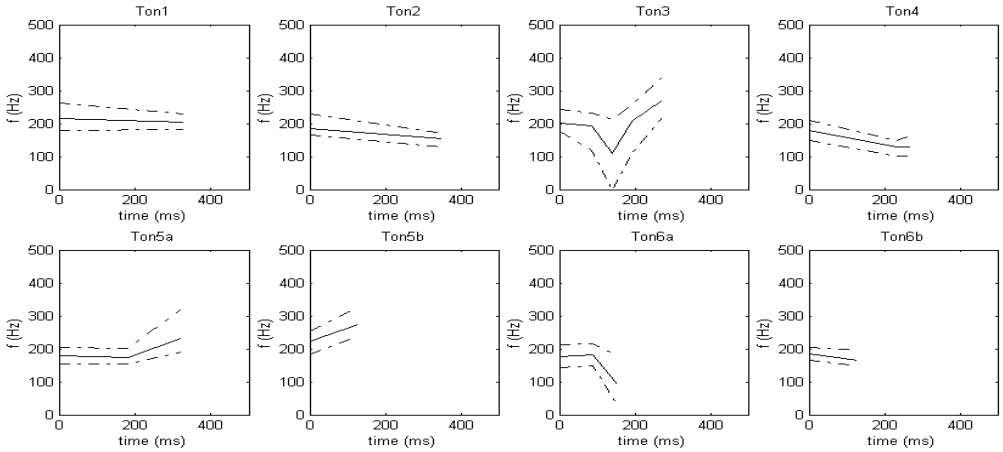


Figure A.3 Gabarits des six tons du sujet féminin DPQ du Nord

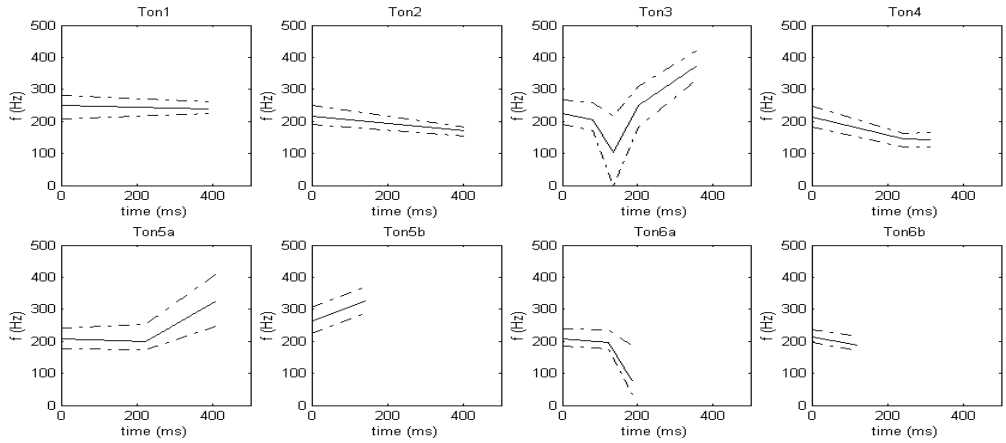


Figure A.4 Gabarits des six tons du sujet féminin DHH du Nord

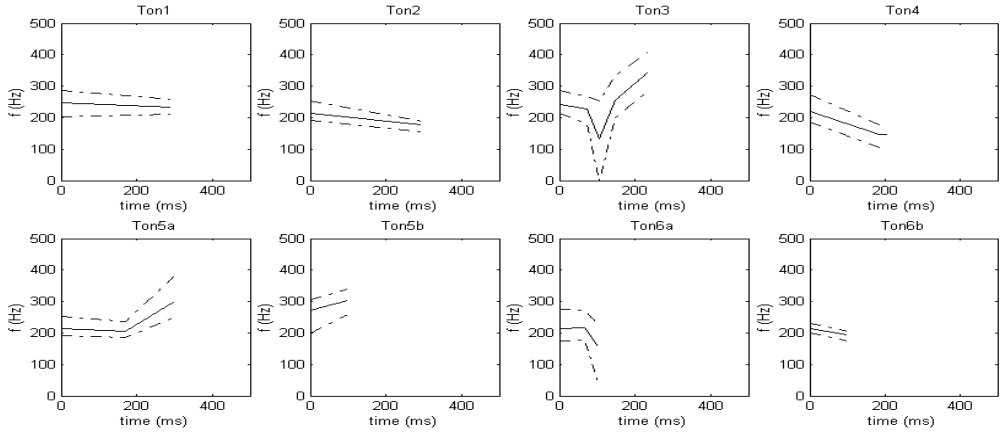


Figure A.5 Gabarits des six tons du sujet féminin DHL du Nord

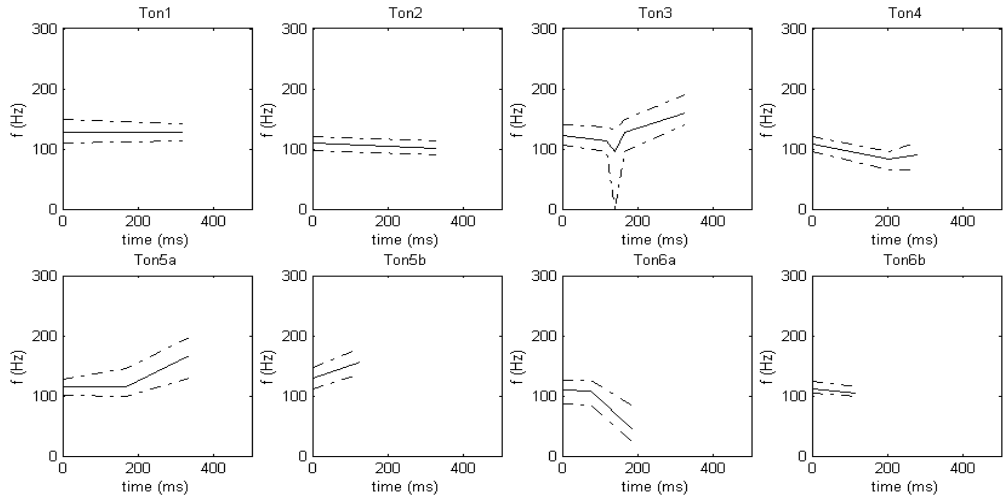


Figure A.6 Gabarits des six tons du sujet masculin BXH du Nord

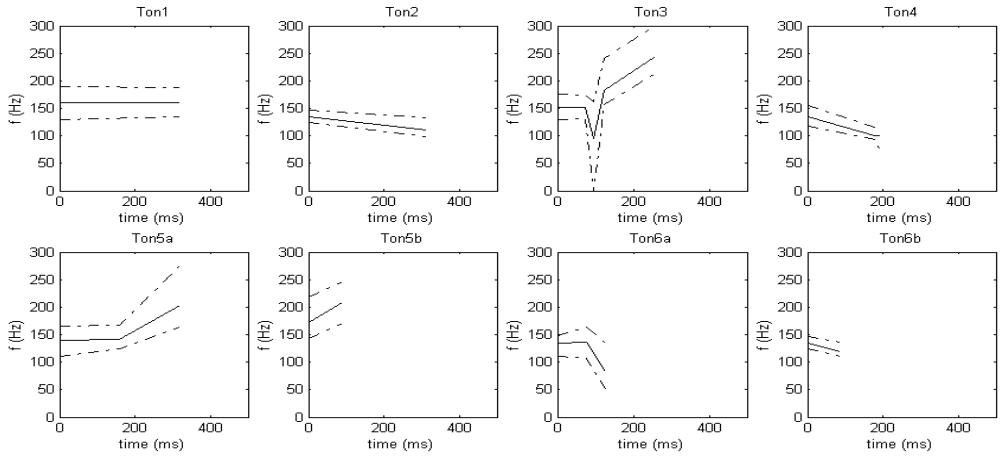


Figure A.7 Gabarits des six tons du sujet masculin TTA du Nord

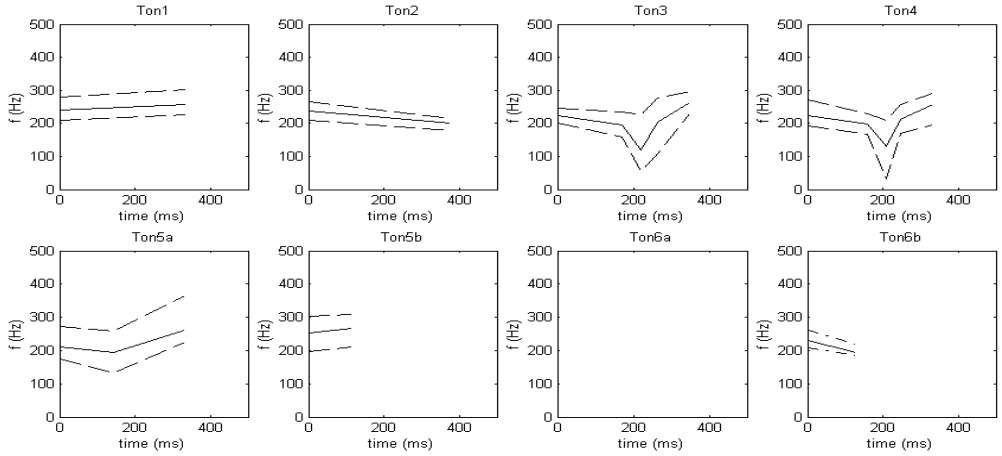


Figure A.8 Gabarits des six tons du sujet féminin NTH du Centre

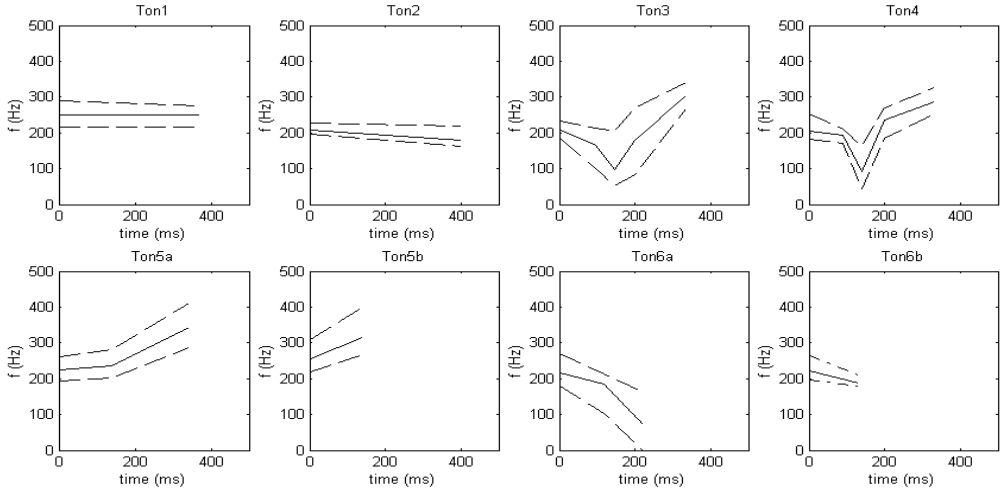


Figure A.9 Gabarits des six tons du sujet féminin VTH du Centre

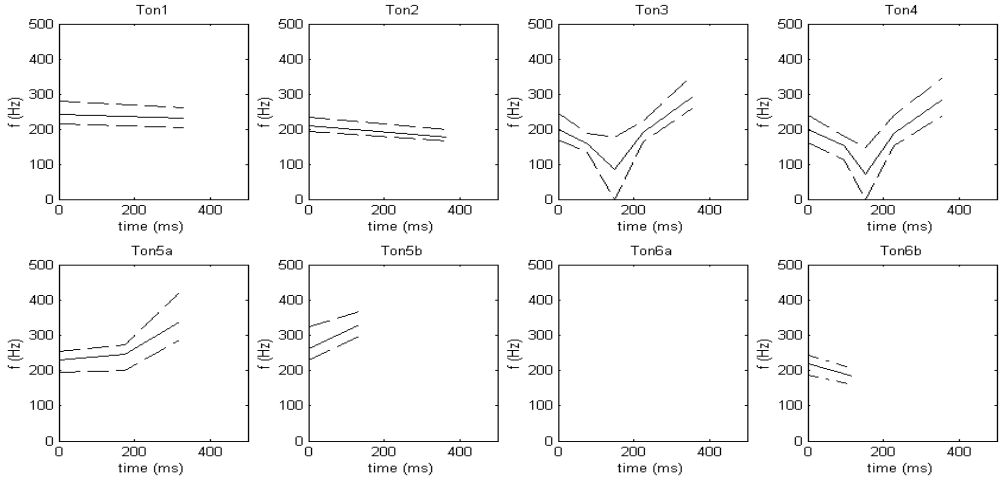


Figure A.10 Gabarits des six tons du sujet féminin BKH du Sud

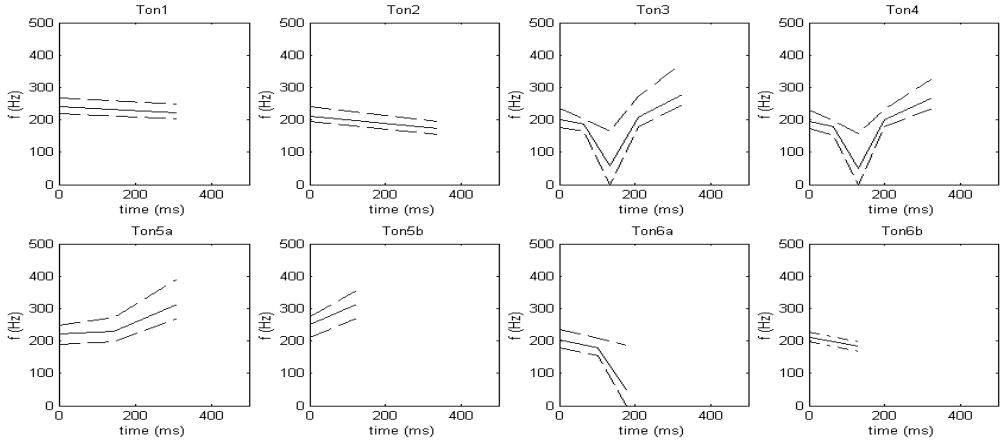


Figure A.11 Gabarits des six tons du sujet féminin LPL du Sud

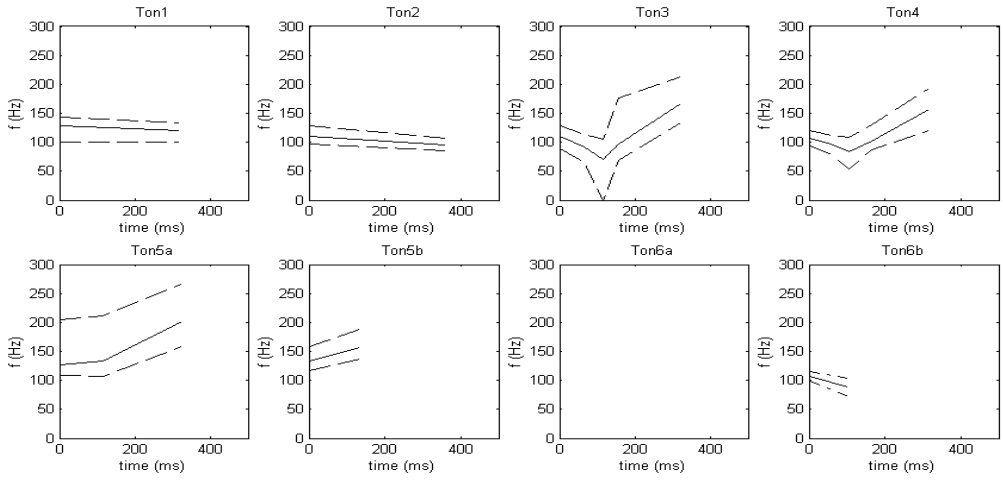


Figure A.12 Gabarits des six tons du sujet masculin LVS du Centre

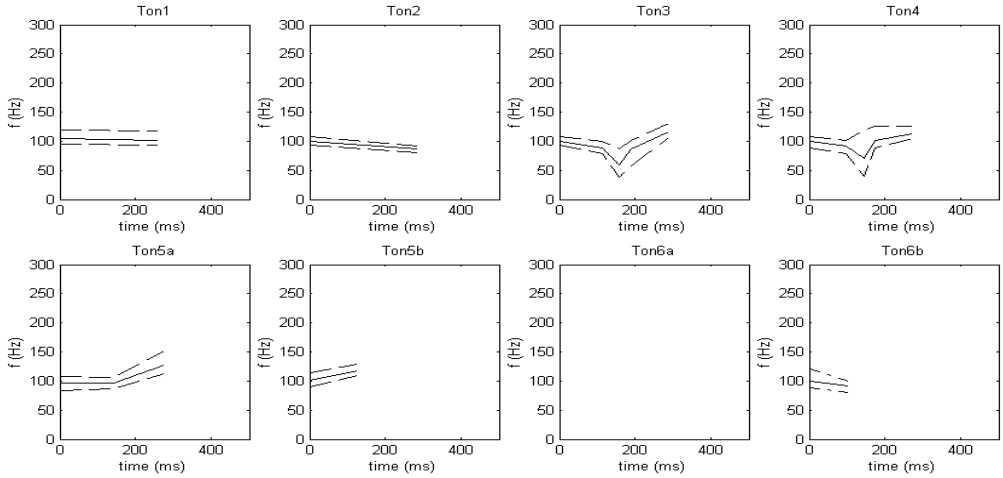


Figure A.13 Gabarits des six tons du sujet masculin TVH du Centre

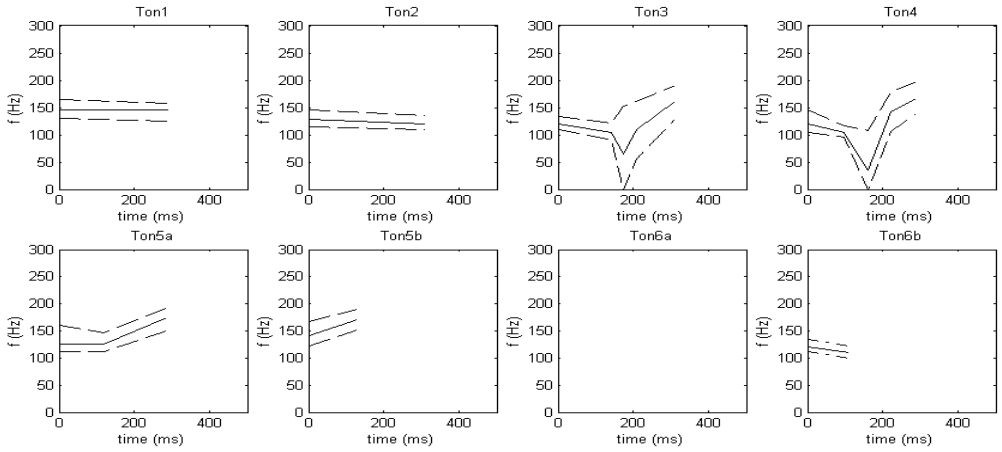


Figure A.14 Gabarits des six tons du sujet masculin HBQ du Sud

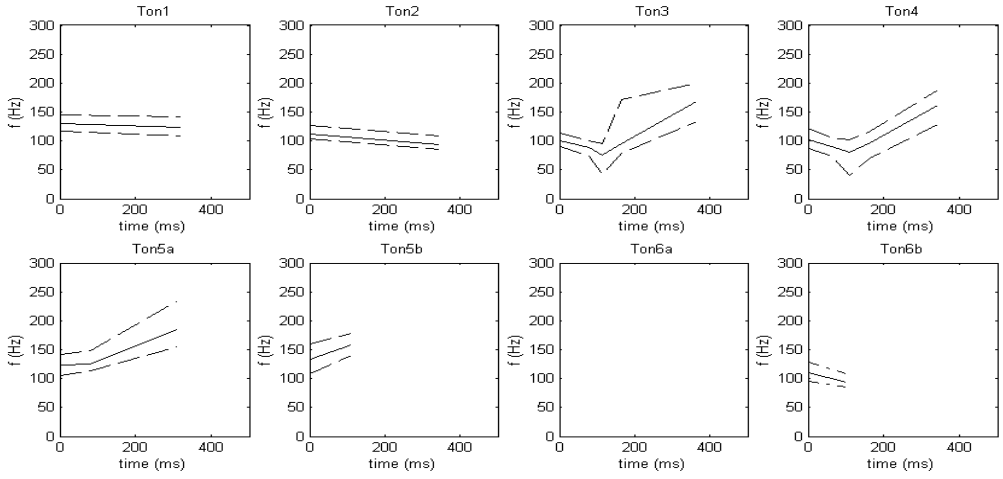


Figure A.15 Gabarits des six tons du sujet masculin TTT du Sud

Annexe B

Rehaussement de la parole

B.1 Introduction

Un problème particulièrement fréquent en traitement du signal est la réduction de bruit. Le résoudre lors de la prise de son de parole pour des applications de téléphonie ou de dialogue homme-machine consiste, en effet, à s'affranchir des conditions environnementales. Un signal de parole enregistré par des microphones dans une situation réelle est normalement dégradé par les perturbations ambiantes, qui peut être éventuellement un signal de parole provenant d'un autre locuteur. A la sortie de chaque microphone, le signal reçu est un composite de sources inconnues, selon un mélange inconnu, qui constitue le signal enregistré dégradé. Avant l'exploitation du signal, un pré-traitement est souvent nécessaire : c'est le rehaussement de la parole.

Les méthodes de rehaussement de la parole sont nombreuses. Elles se divisent en plusieurs catégories suivant le nombre de senseurs dont dispose le système et suivant le nombre de sources sonores.

Une méthode de rehaussement du signal de parole consiste à utiliser la technique dite de séparation de sources. Nous allons présenter en résumé le principe de séparation des sources pour le rehaussement du signal de parole par l'algorithme de séparation de sources fondé sur les cumulants croisés d'ordre 4 [Nguyen Thi,1993].

B.2 Rehaussement de la parole par la séparation de sources dans un mélange convolutif

B.2.1 Modèle de mélange convolutif

Soit deux microphones pour capter les signaux de deux sources, à la sortie des microphones, on observe une superposition des signaux primitifs inconnus selon un mélange inconnu. En général, c'est un mélange convolutif des signaux à large bande, qui dépend de la propagation des signaux dans le milieu, de la position des microphones et des sources, et des caractéristiques de la salle. Ce modèle a été suggéré par [Feder et al, 1989]. Les équations en z des signaux du mélange s'écrivent alors:

$$\begin{aligned} Y_1 &= A_{11}(z).X_1(z) + A_{12}(z).X_2(z) + B_1(z) \\ Y_2 &= A_{21}(z).X_1(z) + A_{22}(z).X_2(z) + B_2(z) \end{aligned} \qquad \text{(B.1)}$$

où: X_1 et X_2 sont deux sources inconnues supposées indépendantes, Y_1 et Y_2 sont les deux signaux observés à la sortie des microphones, B_1 et B_2 sont des bruits additifs qui représente des erreurs de mesure, $A_{ij}(z)$ sont les fonctions de transfert des filtres linéaires avec $i,j \in [1,2]$. La figure **B.1** présente ce modèle du mélange.

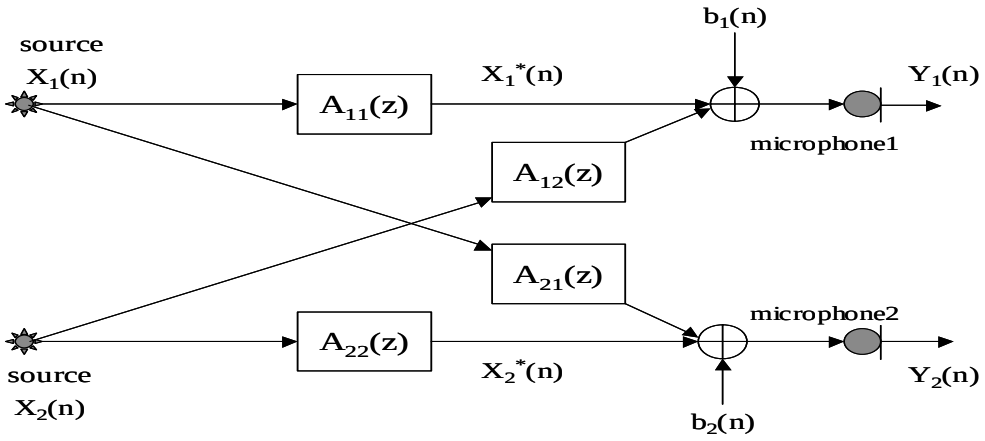


Figure B.1 Modèle général du mélange convolutif

Pour le rehaussement de la parole, un modèle simplifié du mélange convolutif est proposé [Nguyen et al, 1994]. Dans lequel, les filtres A_{ij} sont les filtres à réponse impulsionnelle finie (FIR) causaux, un microphone est placé près du locuteur et l'autre près de la source de

perturbation. On peut ainsi voir les filtres A_{11} et A_{22} comme des scalaires égaux à 1 et les erreurs $b_i(n)$ négligeables. L'équation du mélange (B.1) devient:

$$\begin{aligned} Y_1 &= X_1(z) + A_{12}(z).X_2(z) \\ Y_2 &= X_2(z) + A_{21}(z).X_2(z) \end{aligned} \qquad \text{(B.2)}$$

En supposant les filtres avec l'ordre M, l'équation du mélange simplifié à l'instant discret *n* s'écrit:

$$Y(n)_i = X_i(n) - \sum_{k=0}^{M} a_{ij}(k) X_j(n-k) \qquad \text{avec } i \neq j \quad \text{et} \quad i, j \in [1,2] \qquad \text{(B.3)}$$

où les sources $X_i(n)$ et les filtres A_{ij} sont toujours inconnus.

B.2.2 Architecture et critère de séparation de sources

B.2.2.1 Architecture récursive

Une solution de la séparation de sources des signaux à large bande dans le cas du mélange convolutif fondée sur l'architecture récursive de Herault-Jutten [Jutten et al, 1991] est proposé par [Nguyen et al, 1995]. La figure **B.2** présente de cette architecture. Pour laquelle, les sorties $S_i(n)$ s'écrivent alors:

$$\begin{aligned} S_i(n) &= Y_i(n) - \sum_{k=0}^{M} c_{ij}(k) S_j(n-k) \qquad \text{avec } i \neq j \quad \text{et} \quad i, j \in [1,2] \\ S_i(z) &= Y_i(z) - C_{ij}(z) S_j(z) \end{aligned} \qquad \text{(B.4)}$$

avec $c_{ij}(k)$ est le *k*ème coefficient du filtre C_{ij} et M est l'ordre du filtre.

et encore:

$$S_i(z) = \frac{\left(1 - C_{ij}(z).A_{ji}(z)\right) X_i(z) + \left(A_{ij}(z) - C_{ij}(z)\right).X_j(z)}{\left(1 - C_{ij}(z).C_{ji}(z)\right)} \qquad \text{(B.5)}$$

Il y a deux solutions pour séparer des signaux:

$C_{ij}(z) = A_{ij}(z)$ pour lequel $S_i = X_i$

$C_{ij}(z) = 1/A_{ij}(z)$ pour lequel $S_i = A_{ij}(z)X_j$

Si l'on impose aux $C_{ij}(z)$ d'être des filtres à réponse impulsionnelle finie, seule la première solution est viable (les $A_{ij}(z)$ sont supposés être des filtres à réponse impulsionnelle finie).

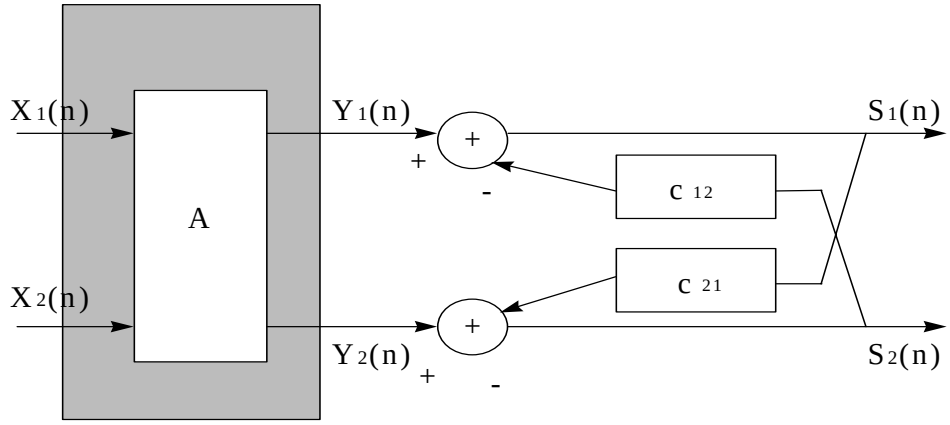


Figure B.2 Architecture de séparation de sources (après Nguyen et al, 1995]

Lorsque les sources X_i et les filtres A_{ij} sont inconnus, alors on ne peut pas trouver directement les filtres C_{ij}. Les coefficients des filtres sont ajustés par un algorithme adaptatif pour rendre les sorties indépendantes. Celui peut être fondé sur l'algorithme d'itération stochastique. Dans lequel, à chaque instant *n*, les kèmes coefficients des filtres C_{ij} sont ajustés en annulant un critère $\phi_{ij}(n,k)$ selon la règle suivante:

$$c_{ij}(n+1,k) = c_{ij}(n,k) - \mu_{ij}\phi_{ij}(n,k) \qquad \text{(B.6)}$$

où: μ_{ij} est le pas d'adaptation, $c_{ij}(n,k)$ est le $k^{ème}$ coefficient du filtre à l'instant *n* .

Si on suppose que les sources X_i sont indépendantes statiquement, et si la séparation est réussie, alors les sorties S_i estimées deviennent indépendantes statistiquement. Le critère $\phi_{ij}(n,k)$ est fondé donc sur l'indépendance statistique des S_i.

B.2.2.2 Algorithme fondé sur les cumulants croisés d'ordre 4

Les deux signaux sont indépendants rigoureux si et seulement si tous les cumulants croisés de tous ordres sont nuls. Mais celui n'est pas soluble, car il nécessite un nombre infini des équations. On peut donc considérer que les deux signaux S_i et S_j sont indépendants, si les deux cumulants croisés d'ordre 4 sont nuls [Nguyen et al, 1995].

$$Cum_{31}(i,j,n,k) = E\left\{s_i^3(n).s_j(n-k)\right\} - 3.E\left\{s_i^2(n)\right\}.E\left\{s_i(n).s_j(n-k)\right\} = 0$$
$$Cum_{31}(j,i,n,k) = E\left\{s_i(n-k).s_j^3(n)\right\} - 3.E\left\{s_j^2(n)\right\}.E\left\{s_i(n-k).s_j(n)\right\} = 0 \qquad \text{(B.7)}$$
$$\forall k \in [0,M], \quad \text{M est l'ordre des filtres}$$

Ici s_i est s_j sont les signaux centrés des S_i et S_j.

E{.} est moment croisé. Le moment croisé peut être calculé des différentes façons:

- la moyenne sur une de N échantillons:

$$E\left\{s_i^l(n)s_j^m(n)\right\} = \frac{1}{N}\sum_{k=n}^{n+N-1} s_i^l(k)s_j^m(k) \qquad \text{(B.8)}$$

- une estimation de la moyenne par filtrage passe-bas du premier ordre

$$E\left\{s_i^l(n)s_j^m(n)\right\} = E\left\{s_i^l(n-1)s_j^m(n-1)\right\} + \frac{s_i^l(n)s_j^m(n) - E\left\{s_i^l(n-1)s_j^m(n-1)\right\}}{T} \qquad \text{(B.9)}$$

où T est une constant à choisir.

En effet, la deuxième méthode est utilisée essentiellement pour implémenter les algorithmes en direct. Car il n'a pas besoin de mémoriser les échantillons futurs et il estime la moyenne en façon lisse.

Le principe de l'algorithme adaptatif de séparation de sources vise à annuler les cumulants croisés d'ordre 4 $S_i(n)$ et $S_j(n-k)$ pour $0 \leq k \leq M$. Les coefficients des filtres estimés sont ajustés en annulant les critères fondés sur les cumulants croisés. Expérimentalement, on observe que les cumulants peuvent s'annuler avec une pente qui peut être positive ou négative [Nguyen Thi et al, 1994]. La règle d'adaptation est proposée donc:

$$c_{ij}(n+1,k) = c_{ij}(n,k) - \mu \left(\frac{\partial Cum_{31}(i,j,n,k)}{\partial c_{ij}(n,k)}\right) Cum_{31}\left(s_i(n), s_j(n-k)\right) \qquad \text{(B.10)}$$

avec $\mu > 0$

B.2.2.3 Pas d'adaptation

En général, les pas d'adaptation des coefficients différents peuvent être différents et variables au cours du temps. Une estimation a été proposée [Nguyen et al, 1996] pour normaliser les

pas d'adaptation des filtres dans les équations (B.10) qui sont fondées sur les moments croisés et les cumulants croisés d'ordre 4:

$$\mu_{ij}(n) = \frac{\mu}{P_{S_i}(n)} \qquad \text{avec } P_{S_i}(n) = \frac{1}{(M+1)} \sum_{k=0}^{M} S_i^4(n-k) \tag{B.11}$$

B.2.2.4 Apprentissage non permanent

Pour les signaux non-stationnaires comme le signal de parole, les apprentissages non-permanents sont utilisés. Pendant les moments de silence de parole, l'énergie du signal est théoriquement nulle, et les variations des coefficients $c_{ij}(n)$ sont voisines de zéro. C'est à dire que l'algorithme n'apprend plus. Un apprentissage permanent n'est donc pas nécessaire. Une méthode d'apprentissage non-permanent a été proposée par [Nguyen et al, 1996] qui est fondée sur l'énergie des sources estimées. A chaque instant *n* on calcule l'énergie des signaux estimés $Es_i(n)$ et $Es_j(n)$:

- SI l'énergie $Es_i > \theta$ et l'énergie $Es_j > \theta$ ALORS deux filtres sont ajustés
- SI l'énergie $Es_i < \theta$ et l'énergie $Es_j > \theta$ ALORS seul le filtre C_{ij} est ajusté
- SI l'énergie $Es_i > \theta$ et l'énergie $Es_j < \theta$ ALORS seul le filtre C_{ji} est ajusté
- SI l'énergie $Es_i < \theta$ et l'énergie $Es_j < \theta$ ALORS les deux filtres ne sont pas ajustés

B.3 Mesures des performances

B.3.1 Mesures des performances en cas simulé

Dans les cas simulés, on connaît exactement les filtres du mélange A_{ij}, les sources. Les signaux mélangés peuvent être calculés numériquement par l'équation (B.2).

- La performance de rehaussement de la parole est mesurée par le rapport du signal sur bruit:

$$RSB_{Si} = 10 \log \left(\frac{\sum_{n=0}^{N-1} X_i^2(n)}{\sum_{n=0}^{N-1} \left(S_i(n) - X_i(n) \right)^2} \right) \tag{B.12}$$

avec X_i est la source, S_i est le signal estimé de la source X_i et N est le nombre des échantillons d'une trame.

- La performance de l'algorithme peut être estimée par les erreurs paramétriques quadratiques. A l'instant n, ces erreurs sont calculées par:

$$QErr\_pc_{ij} = \sum_{k=0}^{M} \left(c_{ij}(n,k) - a_{ij}(n,k) \right)^2 \qquad \text{(B.13)}$$

où M est l'ordre des filtres.

- Les erreurs paramétriques quadratiques moyennes dans une trame sont:

$$QErr\_c_{ij} = \frac{1}{N} \sum_{n=0}^{N-1} QErr\_pc_{ij}(n) \qquad \text{(B.14)}$$

B.3.2 Mesures des performances dans les cas réels

Dans les cas réels, on ne connaît pas les sources X_i et les filtres mélangés A_{ij}. C'est pourquoi on ne peut pas mesurer les rapports du signal sur bruit et ainsi que les erreurs paramétriques quadratiques. Dans ce cas, la convergence des filtres C_{ij} peut être vérifiée par la somme quadratique $QS\_c_{ij}(n)$ des coefficients $c_{ij}(n,k)$, à chaque instant n:

$$QSc_{ij}(n) = \sum_{k=0}^{M} c_{ij}^2(n,k) \qquad \text{(B.15)}$$

La moyenne de $QSc_{ij}(n)$ pour chaque trame est calculée par:

$$QS\_c_{ij} = \frac{1}{N} \sum_{n=0}^{N-1} QSc_{ij}(n) \qquad \text{(B.16)}$$

En effet, si l'algorithme converge, la somme quadratique moyenne QS_c_{ij} tend approximativement vers une constante pour des mélanges stationnaires.

B.4 Validation de l'algorithme

B.4.1 Mélanges convolutifs simulés

B.4.1.1 Séparation du signal de parole et du bruit blanc

Nous avons appliqué l'algorithme aux mélanges convolutifs simulés dans lesquels la source X_1 est un signal de parole est la source X_2 est un bruit blanc. Dans cette expérience, le signal de parole est extrait à partir du corpus du vietnamien comprenant 6 mots isolés prononcés par sujet PNY. Les filtres du mélange A_{ij} sont:

A_{12} = {0.05903; -0.05268; -0.104256; 0.32734; 0.5997}

A_{21} = {0.04652; -0.018077; 0.03601; 0.280955; 0.46475}

Les signaux du mélange sont Y_1 et Y_2 calculés en utilisant l'équation (B.2).

La figure **B.3** présente le rapport signal sur bruit RSBe du signal bruité Y_1 et le rapport du signal sur bruit RSBs du signal estimé S_1. Une qualité de séparation obtenue dans ce cas: la différence entre RSB du signal estimé et du signal bruité est environ 10dB.

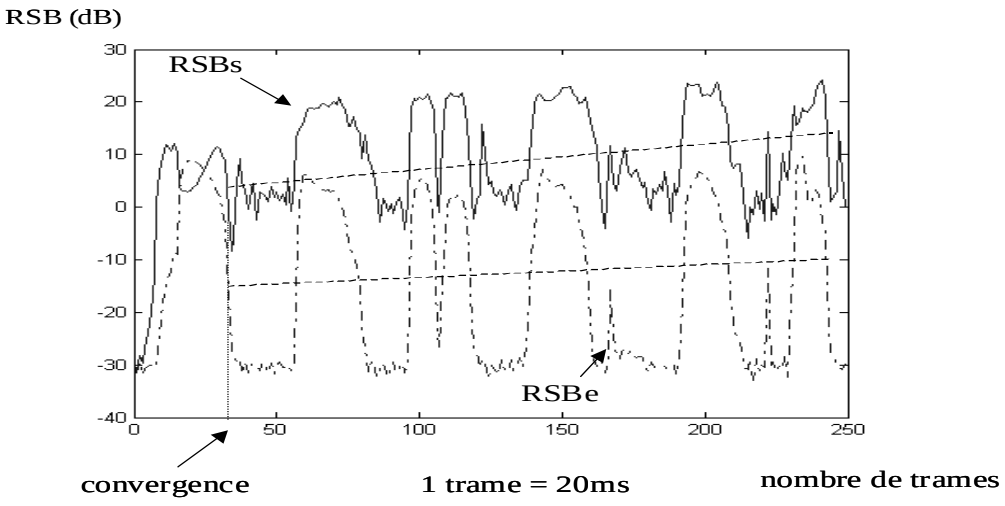


Figure B.3 Les courbes de RSBe du signal bruité et de RSBs du signal estimé dans le cas de séparation du signal de parole et du bruit dans un mélange convolutif simulé

La figure **B.4** présente les courbes des erreurs paramétriques quadratiques moyennes. On peut voir que l'algorithme commence à converger après 30 trames.

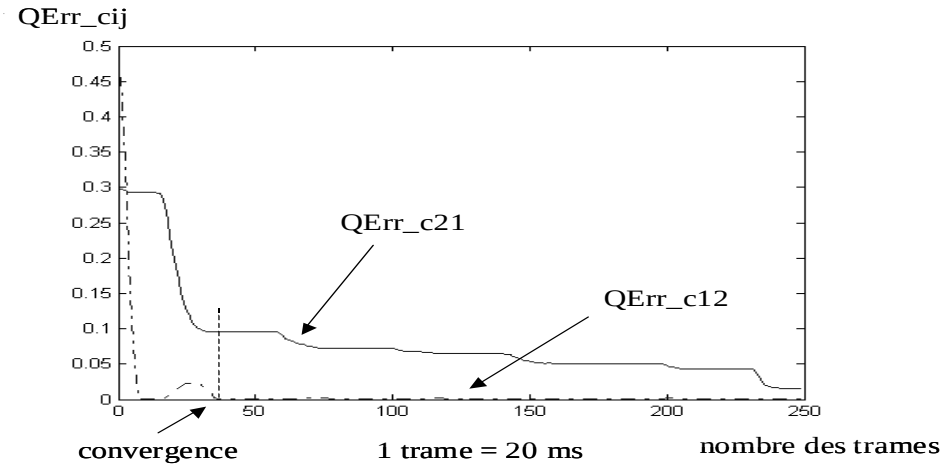


Figure B.4 Les courbes des erreurs paramétriques quadratiques moyennes QErr_cij des filtres estimés C_{ij} par rapport aux filtres A_{ij} dans un mélange convolutif simulé

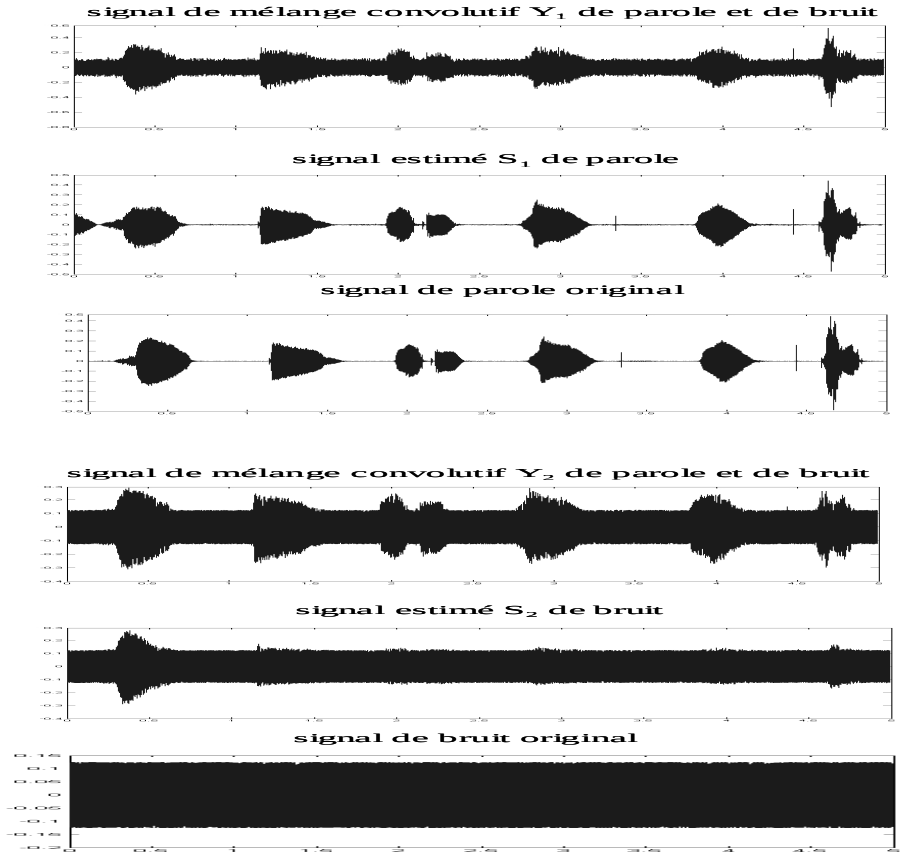


Figure B.5 Séparation de parole et de bruit dans un mélange convolutif simulé

B.4.1.2 Séparation de deux signaux de parole

Dans ce cas, les deux signaux mélangés sont les signaux de parole. Nous utilisons les filtres mélangés comme le cas précédent. Les paroles sont extraites du corpus vietnamien. Le signal de parole X_1 est prononcé par PNY, le X_2 est prononcé par VTT.

La figure **B.6** présente le rapport du signal sur bruit (ici on considère que le signal de parole X_1 est le signal utile, le signal de parole X_2 est le bruit). Dans ce cas-là, la convergence de l'algorithme est un peu plus difficile que celle du bruit blanc (stationnaire). Elle commence à converger après 100 trames. Les signaux de parole sont non-stationnaires. L'adaptation des filtres est lente. En plus il y a deux temps du silence où les coefficients ne sont pas ajustés. C'est ce qu'on peut voir dans la figure **B.7**. Une qualité de séparations obtenue dans ce cas: la différence entre RSB du signal estimé et du signal bruité est environ 10dB d'après de la convergence de l'algorithme.

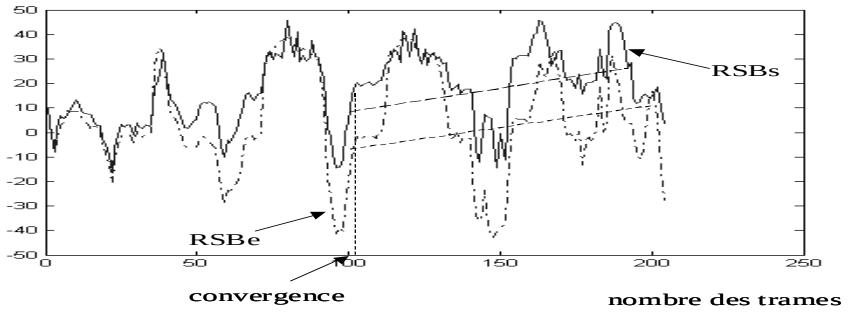


Figure B.6 Les courbes de RSBe du signal bruité et de RSBs du signal estimé dans le cas de séparation de deux signaux de parole dans un mélange convolutif simulé

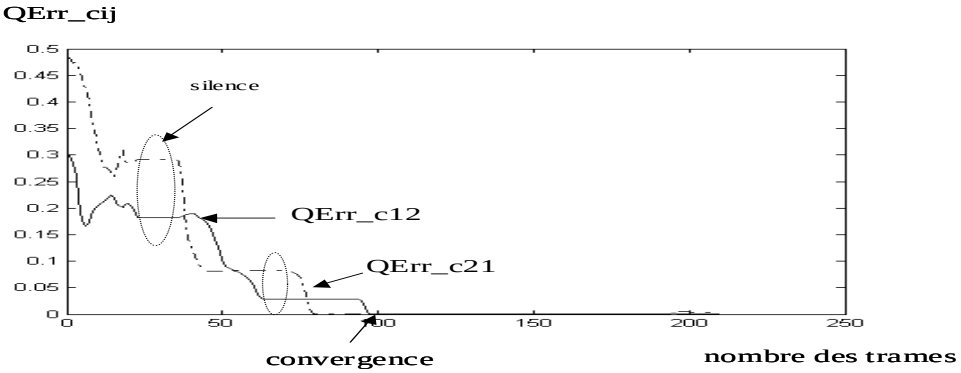


Figure B.7 Les courbes des erreurs paramétriques quadratiques moyennes QErr_cij des filtres estimés C_{ij} par rapport aux filtres A_{ij} dans un mélange convolutif simulé

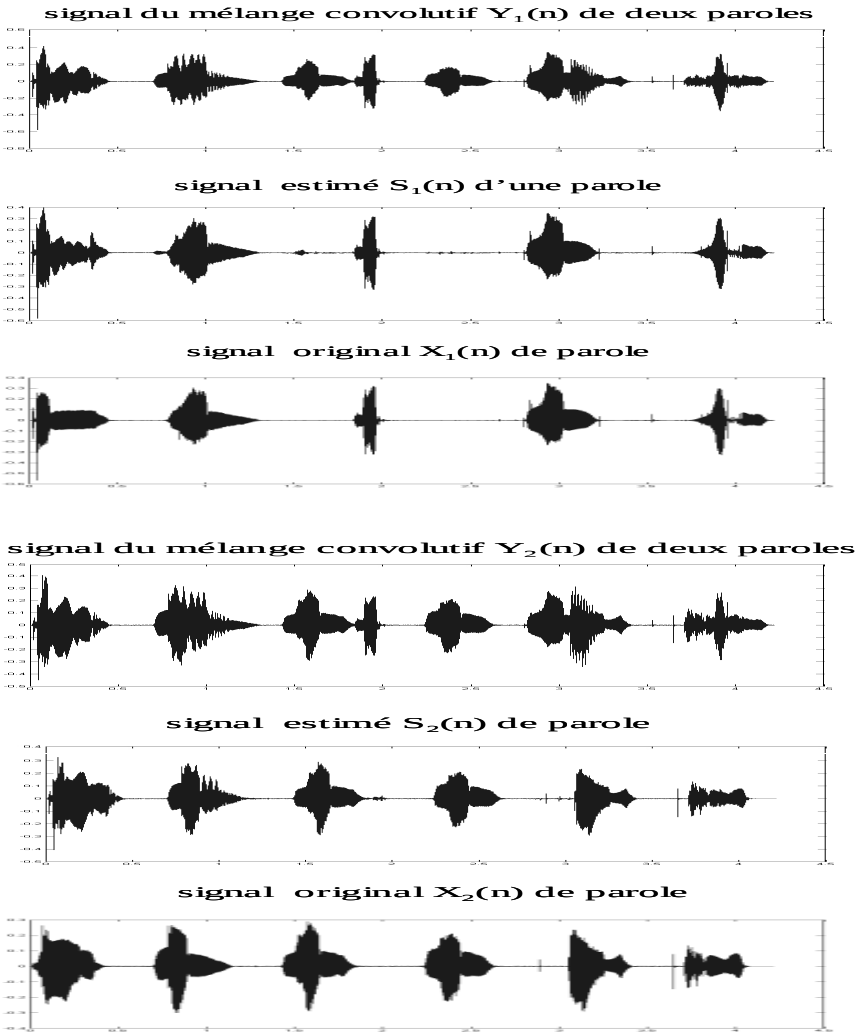


Figure B.8 Séparation de deux paroles dans un mélange convolutif simulé

B.4.2 Expérimentaux réels

Nous avons essayé cet algorithme avec quelques signaux enregistrés réels. On rappelle que dans ce cas où on ne connaît pas les sources et les filtres du mélange. Il y a donc que la moyenne de la somme des erreurs quadratiques qui est calculée.

Nous avons utilisé les signaux enregistrés par Te-Won-Lee [Web01].

- Les signaux enregistrés réels de parole et de musique: Le locuteur prononce les digits de 1 à 10. La cassette joue un morceau de musique. Ces deux sources et les deux microphones sont mis en carré, la distance de chaque côté était de 60 cm.

- Les signaux enregistrés réels de deux extraits de parole: Les locuteurs prononcent les digits de 1 à 10, l'un est en anglais, l'autre est en espagnol. Ces deux sources et les deux microphones sont mis en carré, la distance de chaque côté était 60 cm.

B.4.2.1 Séparation de la parole et de la musique

La figure **B.9** montre les signaux enregistrés du mélange de parole et de musique et les signaux estimés aux sorties. Les mesures de la convergence sont présentés dans la figure **B.10**. Les sommes quadratiques moyennes QS_cij tendent vers des valeurs stables.

Au niveau d'un test auditif, il y a encore le bruit (la musique) dans le signal estimé S_1. Mais on constate une amélioration sensible entre le signal Y_1 du mélange enregistré et le signal S_1 à la sortie du modèle.

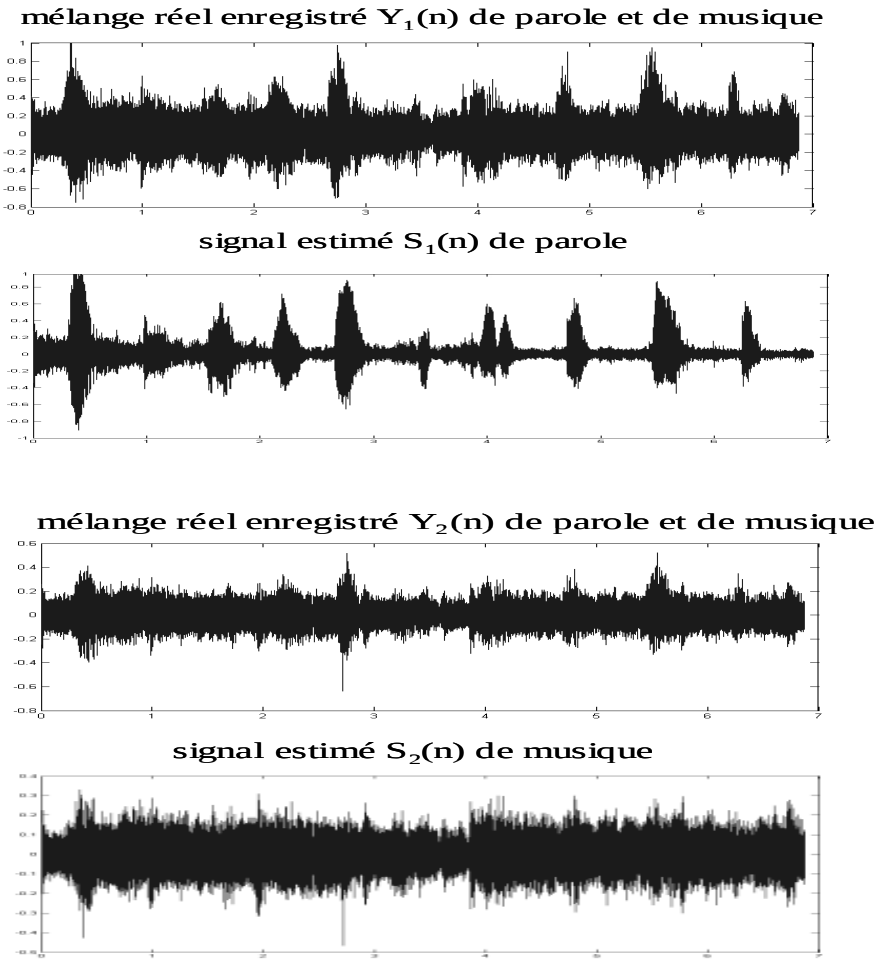


Figure B.9 Séparation de parole et de bruit réel (musique) dans un mélange réel enregistré

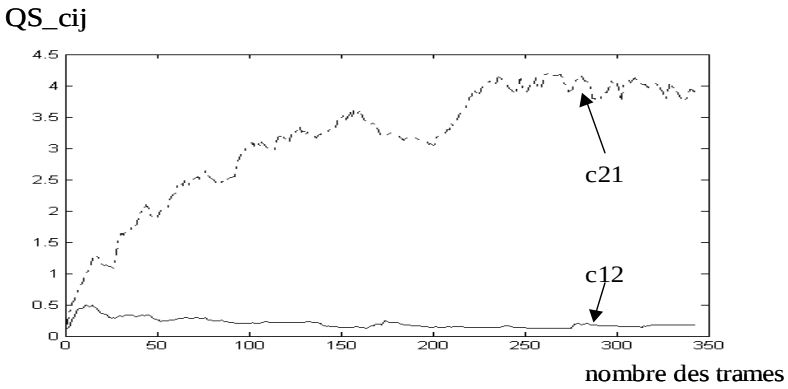


Figure B.10. Les courbes QS_cij des coefficients des filtres estimés en cas réel, obtenues dans la séparation de parole et de bruit

B.4.2.2 Séparation de deux parole

La figure **B.11** montre les signaux enregistrés et les signaux estimés de deux paroles. Les sommes quadratiques moyennes QS_cij tendent vers des valeus stables (figure **B.12**).

A l'écoute, on perçoit déjà une séparation assez nette de deux signaux de parole.

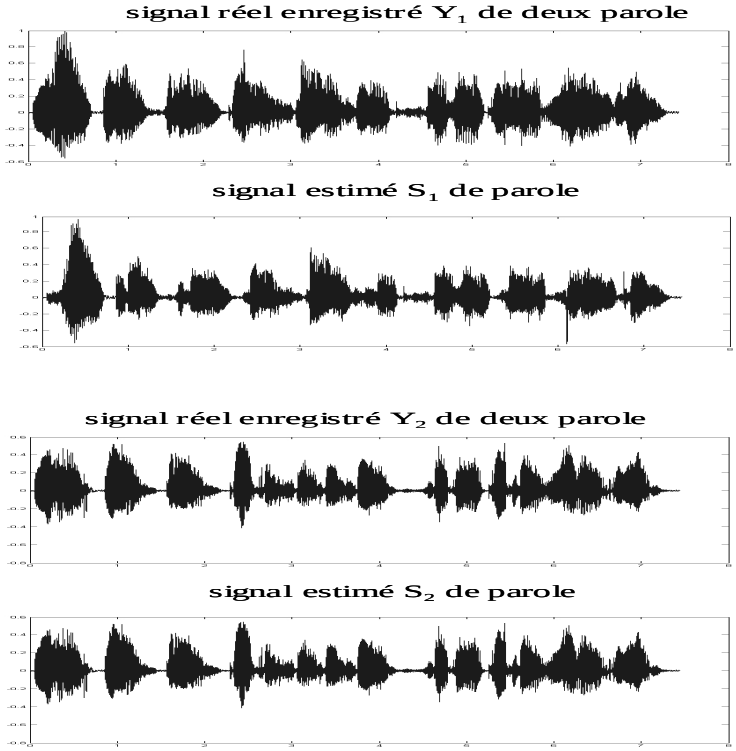


Figure B.11 Séparation de deux paroles dans un mélange réel enregistré

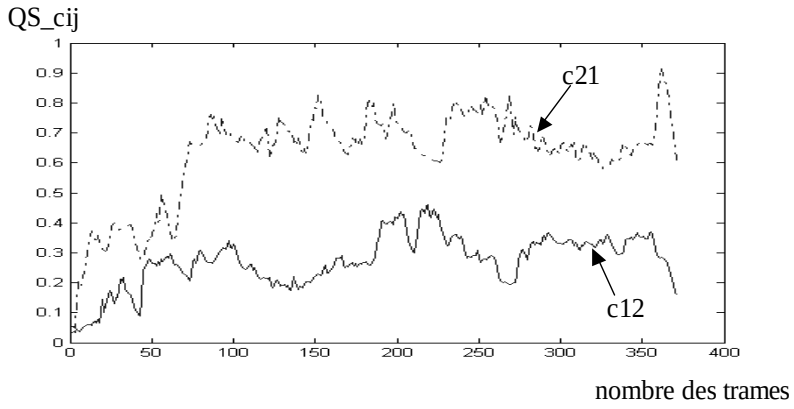


Figure B.12 Les courbes QS_cij des coefficients des filtres estimés en cas réel, obtenues dans la séparation de deux paroles

B.5 Implémentation sur DSP TMS320C6071

Les calculs d'un système de reconnaissance de la parole deviennent de plus en plus importants. Le système doit traiter plusieurs tâches pour la reconnaissance acoustique, la recherche N-meilleurs et l'intégration du modèle de langage, etc. Les calculs de l'algorithme de séparation de sources sont aussi lourds. En réalité, on préfère réaliser la séparation dans un système embarqué, ce qui permet aussi de l'utiliser dans d'autres applications, comme la téléphonie. Dans un système de reconnaissance de la parole complet, il sera alors un système de pré-traitement du signal qui est indépendant du système de reconnaissance. Pour ce but, nous avons implanté la séparation de sources sur la carte DSP de Texas Instruments TMS320C6701. La carte TMS320C6071 contient les caractéristiques suivantes :

- Un DSP TMS320C6701
  - DSP à virgule flottante
  - Performance jusqu'à 1 GFLOPS (giga floating-point opérations per second) à une fréquence d'horloge de 167 MHz
  - Jusqu'à 8 instructions de 32 bits possible par cycle
  - 32 registres de 32 bits
  - 8 unités fonctionnelles: 6 ALUS, 2 multiplieurs
  - Un assembleur optimisant qui permet de faire de la programmation en ' C ' sans penser aux opérations en parallèles.
- Interface PCI, avec fonctionnement autonome possible

- CODEC Stéréo 16 bits avec les fréquences d'échantillonnage disponible: de 5.5 kHz à 48kHz

Ce travail a été réalisé en encadrant un stage de fin d'études d'ingénieur à l'Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble réalisé par Barton [Barton J, 2000]. A partir de notre programme en C tournant sur un ordinateur de type PC, Barton a implanté une version embarquée sur la carte TMS320C6071. Sur la carte DSP, l'algorithme fonctionne bien en temps réel avec une fréquence d'échantillonnage de 9600 Hz (suffisante pour la téléphonie) avec un ordre maximum pour les filtres de 20 ou bien avec une fréquence d'échantillonnage plus élevée de 16000 Hz (pour la reconnaissance) mais avec un ordre maximal des filtres de 10.

Les systèmes de reconnaissance de parole utilisés travaillent le plus souvent avec des signaux à une fréquence d'échantillonnage de 16000 Hz. Notre système embarqué fonctionne alors seulement avec des filtres d'ordre 10. Cet ordre de 10 est encore bas pour que la séparation de sources en réel soit efficace. Pour augmenter cet ordre, tout en continuant à tourner en temps réel, une optimisation du code du programme sur DSP sera nécessaire.

B.6 Conclusions

Nous avons présenté une solution de rehaussement du signal de parole fondée sur la séparation de sources en utilisant les cumulants croisés d'ordre 4. Cette méthode a été proposée par Nguyen et Jutten [Nguyen et al, 1995]. Cette première étude a été continué par les travaux de Charkani pour l'application à la téléphonie mains-libres dans les voitures [Charkani, 1996] et de Taloud [Taloud, 1997] pour le rehaussement de la parole dans la reconnaissance en environnements bruités.

De notre côté, nous avons implanté cet algorithme sur une carte à base de processeur spécialisé pour réaliser un module autonome et fonctionnant en temps réel. Les résultats que nous avons obtenus contribuent à vérifier son efficacité. Cependant, il reste encore à optimiser l'algorithme, en particulier la gestion des ressources mémoires sur la carte DSP pour pouvoir intégrer cette méthode dans le système de reconnaissance comme une partie de pré-traitement du signal, avec un taux de rehaussement acceptable.

B.7 Références

Charkani N, 1996
Séparation Auto-adaptative de sources pour des mélanges convolutifs. Application à la téléphonie mains-libres dans les voitures
Thèse, INP Grenoble

Feder M, Oppenheim A.V, Weinstein E, 1989
Maximum likelihood noiise cancellation using the EM algorithm
IEEE Trans on ASSP, Vol 37, No 2, 204-216

Jutten C, Herault J, Comon P, Sorouchyari E, 1991
Blind separation of sources: Part I, Part II, Part III
Signal Processing, vol 24, pp 1-29

Nguyen Thi H-L, 1993
Séparation aveugle de sources à bande large dans un mélange convolutif, application au rehaussement de la parole
Thèse, INP Grenoble

Nguyen Thi H-L, Jutten C, 1995
Blind Source Separation for Convolutive Mixtures
Signal Processing, Vol 45, pp 209-229

Nguyen Thi H-L, Jutten. C, Kabré. H, Vaelen. J, 1996
Separation of Sources: A Method for Speech Enhancement
Applied Signal Processing, Vol 3, 177-190

Taloud P-Y, 1997
Rehaussement de la parole pour améliorer la reconnaissance en environnements bruités
D.E.A, ICP/INP de Grenoble

Texas Instruments, 1998
TMS320C6201/6701 Evaluation Module User's Guide

Texas Instruments, 1999
TMS320C6000 Code Composer Studio Tutorial

Web01:

http://www.cnl.salk.edu/~tewon/ica_cnl.html

Annexe C

Liste des publications de l'auteur

Pham Thi N.Y, Castelli E, Nguyen Q.C (2002)
Gabarits des tons vietnamiens
JEP2002, Nancy (France), accepté

Nguyen Q.C, Castelli E (2002)
Caractérisation et Reconnaissance automatique des tons du vietnamien
RFIA2002, volume 2, pp 529-537, Angers (France)

Nguyen Q.C, Pham N.Y, Castelli E (2001)
Shape Vector Characterization of Vietnamese Tones and Application to Automatic Recognition
ASRU2001, Madonna di Campiglio (Italy)

Nguyen Q.C, Pham N.Y, Castelli E (2001)
Premier moteur de reconnaissance automatique du vietnamien en mots isolés
Conférence de l'IPH, Hanoi (Vietnam)

Nguyen Q.C, Istrate D, Barton J & Castelli E (2001)
Blind Source separation
Student Forum – présentation Poster - ICASSP2001

RECONNAISSANCE DE LA PAROLE EN LANGUE VIETNAMIENNE

Résumé

La langue vietnamienne est une langue asiatique tonale avec six tons lexicaux dans laquelle chaque syllabe est prononcée avec un ton lexical distinct. Le principal objectif de ce travail de thèse est la réalisation d'un moteur de reconnaissance automatique de la parole en langue vietnamienne. Dans les langues tonales, pour reconnaître les mots, il faut réaliser la reconnaissance des syllabes prononcées avec le ton, c'est-à-dire réaliser deux processus parallèles de reconnaissance du ton et de reconnaissance de la syllabe indépendamment du ton. Nous avons réalisé un corpus multi-locuteurs qui constitue une base de données en langue vietnamienne pour la caractérisation des six tons du vietnamien. Les gabarits des tons sont alors déduits pour permettre de définir un vecteur caractéristique nécessaire à la reconnaissance des tons vietnamiens. Notre système de reconnaissance des tons vietnamiens utilise la technique fondée sur les modèles de Markov cachés (HMM). Le taux de reconnaissance des tons atteint est de 93.3% en mode dépendant du locuteur et de 91.5% en mode indépendant du locuteur. Un module de reconnaissance de la syllabe vietnamienne indépendamment du ton, lui aussi fondé sur l'utilisation de HMMs est alors couplé au module de reconnaissance des tons pour réaliser un prototype de reconnaissance du vietnamien en mots isolés. Le taux de reconnaissance du moteur complet est de 80.7% en mode de mots isolés.

Mots clés: reconnaissance de la parole, caractérisation, ton, syllabe, vietnamien, modèle de Markov caché, corpus.

Abstract

The Vietnamese language is a tonal Asian language with six lexical tones in which each syllable is pronounced with a distinct lexical tone. The main goal of this thesis is to realize a system of automatic speech recognition in Vietnamese language. In the tonal languages, in order to recognize the words, it is necessary to perform the recognition of the syllables pronounced with the tone, i.e., to perform two parallel processes: the recognition of the tone and the recognition of the base-syllable (i.e., the syllable disregarding the tone). We carried out a multi-speakers corpus that constitutes a database in Vietnamese language for the characterization of the six tones of Vietnamese. The shapes of the tones are then deduced to make it possible to define a characteristic vector for Vietnamese tone recognition. Our tone recognition system uses the technique based on the hidden Markov model (HMM). The tone recognition rate achieved 93.3% in speaker dependent mode and 91.5% in speaker independent mode. A Vietnamese base-syllable recognition module also based on HMMs, is then coupled with the tone recognition module to carry out a prototype of Vietnamese recognition in isolated words. The rate of recognition of the complete system achieved 80.7% in isolated words mode.

Key words: speech recognition, characterization, tone, syllable, Vietnamese, hidden Markov model, corpus.