Prise en compte de critères acoustiques pour la synthèse de la parole

Soufiane Rouibia

N° d'ordre: 2006tel0020

THÈSE

présentée devant

L'ECOLE NATIONALE SUPERIEURE DES TELECOMMNUNICATIONS DE BRETAGNE

en habilitation conjointe avec l'Université de Rennes 1

pour obtenir le grade de

DOCTEUR de l'ENST Bretagne

Mention: Traitement du signal et Télécommunications

par

Soufiane ROUIBIA

Prise en compte de critères acoustiques pour la synthèse de la parole

soutenue le 27 Septembre 2006 devant la commission d'examen

Mme. :	R.	Le Bouquin Jeannès	Présidente
MM. :	V.	Aubergé	Rapporteurs
	JF	Bonastre	
MM. :	O.	Boëffard	Examinateurs
	J.M.	Boucher	
	O.	Rosec	

Remerciements

J'ai vécu mes trois années de thèse comme une aventure formidable, jalonnée bien sûr de passages difficiles. Ma plus belle chance a été de travailler avec Monsieur Olivier Rosec, ingénieur de recherche et développement à la division R&D de France Télécom. Il m'est difficile de trouver les mots pour exprimer toute ma reconnaissance envers Olivier Rosec. Je le remercie pour sa très grande disponibilité et pour ses encouragements qui n'ont jamais cessé. Olivier Rosec a guidé mes recherches de façon juste, précise et rigoureuse, tout en m'expliquant chaque facette du métier. Il m'a initié au monde de la recherche et il a fait en sorte que je m'y sente bien. Je suis fier d'avoir été parmi ses étudiants en thèse.

Je remercie vivement Monsieur Jean-Marc Boucher, professeur à l'ENST de Bretagne, qui m'a fait l'honneur d'être mon directeur de thèse. Je le remercie aussi pour le soin qu'il a apporté à la lecture de mon manuscrit, et pour ses remarques et conseils qui ont contribué à son amélioration.

J'aimerais également remercier Madame Régine Le Bouquin Jeannès, Professeur à l'Université de Rennes I, qui m'a fait l'honneur de présider la commission d'examen.

Mes plus sincères remerciements sont adressés à Madame Véronique Aubergé, Chargée de Recherche CNRS à l'Université Stendhal-Grenoble, et à Monsieur Jean-François Bonastre, Maître de Conférence HDR à l'Université d'Avignon et des pays de Vaucluse, pour avoir accepté d'être rapporteur de mes travaux. Leurs critiques constructives m'ont été particulièrement précieuses dans finalisation de ce document. Je remercie également Monsieur Olivier Boëffard, Professeur à l'ENSSAT, d'avoir bien voulu examiner ce travail.

Mes plus sincères remerciements vont également à Monsieur Thierry Moudenc, responsable de l'unité de recherche et développement "Vocalisation Multimodale In-

novante" à France Télécom, qui m'a chaleureusement accueilli dans son équipe. Ses conseils et ses commentaires ont été fort utiles.

Bien sûr, je remercie de tout coeur le petit noyau lannionais de l'équipe VMI de France Télécom R&D qui m'a accueilli et offert un environnement agréable et propice à mon développement tant personnel que professionnel.

Mes remerciements vont également à l'ensemble des personnes qui m'ont accompagné, de près ou de loin, durant ces trois années et qui ont contribué, directement ou non, à l'aboutissement de ce travail. Merci enfin à mes parents et mes frères et sœurs pour leur soutien, à distance mais sans faille, et à Leila pour m'avoir encouragé ou simplement supporté. Sa patience et son attention m'ont été forcément profitables.

À mes parents

Résumé

Cette thèse s'inscrit dans le domaine de la synthèse vocale à partir du texte et traite plus particulièrement de la synthèse par corpus (SPC). Cette approche basée sur la concaténation de segments acoustiques contenus dans de grandes bases de données s'est peu à peu instaurée comme un standard. En effet, moyennant la sélection d'unités adaptées au contexte de synthèse, elle permet d'aboutir à un signal de parole dont le naturel peut être assez bien préservé. La qualité de la synthèse obtenue par la méthode par concaténation est étroitement liée d'une part au corpus de synthèse et d'autre part à l'algorithme de sélection des unités. Malgré le saut notable de qualité qu'a permis d'atteindre cette technologie, la SPC n'est pas capable de garantir une parole dont la qualité soit à peu près constante sur l'ensemble d'un énoncé. Ceci est en grande partie dû au manque de contrôle acoustique des systèmes de SPC actuels. L'objectif de cette thèse est donc d'introduire des mécanismes permettant un meilleur contrôle acoustique lors de la synthèse.

La méthode proposée consiste à effectuer une sélection sur la base d'une cible purement acoustique. Cette cible est déduite de modèles acoustiques - plus précisément des modèles de sénones - estimés lors d'une phase d'apprentissage. Dans un premier temps, nous proposons un algorithme de sélection basé uniquement sur cette cible acoustique. Puis la méthode de sélection est modifiée de manière à mieux contrôler l'information de fréquence fondamentale. Le module de sélection proposé est également combiné à un module de pré-sélection des unités, ce qui conduit à une diminution sensible de la complexité algorithmique sans dégradation perceptible des résultats. Des tests d'écoutes formels révèlent que la méthode proposée permet de réduire significativement les discontinuités acoustiques lors de la concaténation. La méthode proposée est également appliquée à la réduction de corpus acoustiques et conduit à une réduction de l'ordre de 60% de la base acoustique sans dégradation de la qualité de la parole produite.

Abstract

This thesis relates to text-to-speech synthesis and deals more particularly with the corpus based approach. In the last few years, this approach based on the concatenation of acoustic segments contained in large databases has become increasingly popular. Indeed, selecting units which best fit the text to be synthesized leads to a synthesised signal whose naturalness can be rather well preserved.

The quality of the synthesized speech obtained by corpus-based methods is closely related on the one hand to the corpus used for synthesis and on the other hand to the unit selection algorithm. In spite of the notable increase of quality reached with this technology, corpus-based speech synthesis is not able to guarantee a synthesised speech whose quality is constant on an entire utterance. This is mainly due to the lack of acoustic control of the existing corpus-based speech synthesis systems. The main objective of this thesis is therefore to introduce a mechanism allowing a better acoustical control during synthesis.

The proposed method uses statistical approaches to generate a smooth acoustic target from which the sequence of synthesis units will be selected. This target is deduced from acoustic models, namely context dependent senone models, estimated during a training phase. Initially, we propose an algorithm of selection based only on this acoustic target. Then, the proposed selection method is modified so as to better control the information of fundamental frequency. This unit selection module is also combined with a pre-selection module so as to drastically reduce the computational load. Formal listening tests show that the proposed method leads to a significant reduction in acoustic discontinuities during the concatenation. The proposed method is also applied to acoustic database reduction and enables a compression of about 60% of the acoustic database without perceptible decrease of the speech quality.

Table des matières

Table des figures

Liste des tableaux

Introduction

Contexte de la thèse

Un synthétiseur de parole est le résultat d'une imitation particulière et originale de l'acte parlé. Des développements importants dans la synthèse de la parole à partir du texte (Text-To-Speech en anglais) et les techniques de traitement du langage naturel au cours des dernières années ont rendu cette technologie de plus en plus utilisée. Cette avancée est due essentiellement à l'émergence de nouvelles techniques, notamment avec l'utilisation massive des systèmes basés sur la concaténation d'unités acoustiques. Cette dernière consiste à mettre bout à bout des segments de signaux préalablement enregistrés pour générer par la suite un signal de parole synthétique.

Un système de synthèse vocale à partir du texte prend en entrée une forme textuelle et produit en sortie le signal de parole correspondant à une vocalisation de ce texte. Dans un tel système, un premier bloc de traitements linguistiques analyse le texte, détermine la suite phonétique permettant de le vocaliser ainsi que des consignes prosodiques qui spécifient une certaine "mélodie" à restituer lors de la synthèse. Les modules acoustiques génèrent ensuite le signal de synthèse en concaténant des segments de parole naturelle stockés, obtenus à partir de l'enregistrement d'un locuteur professionnel. Les premiers systèmes par concaténation utilisaient des dictionnaires acoustiques de taille réduite (environ 5 Mo par voix) où chaque unité (diphone) a un seul représentant acoustique. A la synthèse, pour respecter les consignes prosodiques prédites, d'importantes déformations sont alors nécessaires. Ces modifications engendrent malheureusement de fortes dégradations de la qualité de la parole synthétisée.

La disponibilité récente de ressources informatiques importantes a permis l'émergence de solutions nouvelles regroupées sous l'appellation de "synthèse par corpus" (SPC).

Dans cette approche, la base de données acoustiques ne se restreint pas à un dictionnaire de diphones monoreprésentés mais contient des unités de taille variable (diphones, triphones, syllabes, etc...) enregistrées dans différents contextes linguistiques (phonétique, syllabique, syntaxique, etc...) et selon différentes variantes prosodiques. La problématique de la synthèse change alors radicalement : il ne s'agit plus de déformer le signal de parole en visant à dégrader le moins possible la qualité du timbre mais plutôt de disposer d'une base de données suffisamment riche et d'une algorithmique fine permettant la sélection des unités de la base les mieux adaptées au contexte. Actuellement, la plupart des systèmes utilisent des corpus dont la taille dépasse 100 Mo. La sélection consiste à déterminer la séquence d'unités ayant les contextes les mieux adaptés et minimisant les discontinuités aux instants de concaténation. L'intérêt de cette approche est que, moyennant une stratégie de sélection adéquate, il devient possible de limiter fortement le recours à un algorithme de modification prosodique. De ce fait, le naturel de la voix peut être préservé. Cette étude s'intéresse plus particulièrement à la sélection des unités dans le cadre de la synthèse par corpus.

Problématique

La synthèse par corpus repose sur le principe "choose the best to modify the least" [Sag88], [BC95], [HB96]. Dans ce type de synthèse [BPQ+99], la recherche de l'unité souhaitée est réalisée sur un corpus qui contient non plus un seul, mais plusieurs représentants de chaque unité, de sorte que les modifications acoustiques à apporter à l'unité sélectionnée soient réduites au strict minimum. Le succès de cette technologie par rapport à la synthèse par diphone tient au fait que moyennant des critères de sélection relativement simples et une base de données suffisamment riche, il devient possible de synthétiser des signaux de parole dont le naturel est assez bien préservé. Ainsi, cette approche s'est standardisée et est devenue la technique la plus utilisée par tous les systèmes de synthèse actuels.

Bien que considérée comme une rupture technologique majeure, de par le saut de qualité qu'elle a pu engendrer, la SPC n'est cependant pas exempte de défauts. En effet, cette technologie ne parvient pas actuellement à garantir une parole synthétique de très haute qualité sur l'ensemble d'un énoncé, ce qui se traduit localement par l'apparition d'artefacts audibles. Ces défauts proviennent soit d'une déficience dans la couverture

acoustique atteinte par le corpus, soit de problèmes liés à la sélection des unités. En tout état de cause, il semble indispensable d'introduire des mécanismes permettant un meilleur contrôle acoustique lors de la synthèse. Des travaux récents dans le domaine de la sélection ont introduit de nouvelles approches [DE98] et [HAH+97], telles que la modélisation statistique du signal, traditionnellement utilisée dans le domaine de la reconnaissance de la parole, pour une étape de présélection. Malheureusement, l'étape de sélection reste basée sur la minimisation de fonctions de coûts ce qui entraîne des discontinuités acoustiques nécessitant un recours à des traitements correctifs. Ainsi, dans ce contexte, l'objectif de cette thèse est de proposer une nouvelle méthode de sélection basée essentiellement sur des critères acoustiques.

Objectifs de la thèse

Cette thèse vise à introduire des critères de cohérence acoustiques dans le processus de sélection des unités afin de limiter les discontinuités acoustiques. Plus précisément, nous définissons des cibles spectrales sur l'ensemble de la phrase à synthétiser. Ces cibles sont déduites de modèles acoustiques estimés lors d'une phase d'apprentissage. Dans un premier temps nous proposons un algorithme de sélection basé uniquement sur cette cible acoustique. Par la suite, nous proposons une amélioration de cet algorithme sur deux aspects : prise en compte de la fréquence fondamentale et réduction de la complexité. Nous appliquons également cette méthode de sélection en vue de réduire les bases acoustiques utilisées par la synthèse.

Organisation du document

Après avoir présenté au chapitre 1 les éléments essentiels relatifs à la synthèse de la parole à partir du texte, nous dressons au chapitre 2 un état de l'art sur la synthèse par corpus. À cette occasion, nous passons en revue les principales méthodes de sélection existantes ainsi que différents critères utilisés en synthèse par corpus.

Le chapitre 3 présente la méthode de sélection des unités acoustiques proposée et les résultats obtenus après différents tests d'évaluation. Dans un premiers temps, nous détaillons la modélisation acoustique qui est effectuée "hors line". Dans la deuxième

partie de ce chapitre, nous présentons les différentes étapes du module de sélection proposé.

Tirant parti des tests menés au chapitre 3, nous proposons au chapitre 4 des perfectionnements de la méthode proposée. La première amélioration consiste en la prise en compte de la fréquence fondamentale tant lors de la classification des unités acoustiques que lors de la synthèse. La deuxième amélioration vise à réduire la complexité de la méthode proposée par l'ajout d'une procédure de présélection en amont de notre module de sélection. Des tests subjectifs et objectifs pour évaluer l'apport de ces améliorations sont présentés à la fin de ce chapitre.

Dans le chapitre 5, nous utilisons le formalisme présenté aus chapitres 3 et 4 à des fins de réduction de bases. La qualité de la synthèse produite à partir des nouvelles bases réduites est évaluée par des tests tant subjectifs qu'objectifs.

Pour finir, une conclusion termine ce document et diverses perspectives de recherche ouvertes par ces travaux sont proposées.

Chapitre 1

Synthèse de la parole

1.1 Introduction

Ce premier chapitre a pour but d'une part d'introduire et de donner un aperçu général sur les différents thèmes abordés dans cette thèse et d'autre part de situer ce travail dans son environnement technique et scientifique. Ainsi, les connaissances essentielles qui décrivent les natures physiologiques et phonétiques de la parole sont d'abord présentées. Ensuite, nous présentons le cadre technique de notre étude : la synthèse de la parole. La présentation s'articule autour des principes de la synthèse vocale, des différents systèmes présents dans l'état de l'art, suivie d'une description bien détaillée de la technique utilisée dans le cadre de cette étude : la synthèse par concaténation d'unités acoustiques.

1.2 Généralités sur la parole

La parole est une faculté, propre à l'homme, de communication par des sons articulés. Elle met en jeu des phénomènes de natures très différentes et peut être analysée de bien des façons. On distingue généralement plusieurs niveaux de description non exclusifs : physiologique, phonologique, phonétique, acoustique, morphologique, syntaxique, sémantique, et pragmatique. Nous survolons dans ce manuscrit les quatre premiers niveaux qui sont les niveaux les plus concernés par notre étude.

1.2.1 Niveau physiologique

Les sons de la parole se produisent lors de la phase d'expiration au cours de laquelle un flux d'air contrôlé, en provenance des poumons passe à travers le larynx et le conduit vocal (conduit respiratoire). Ce flux d'air appelé *air pulmonaire* rencontre sur son passage plusieurs obstacles potentiels qui vont le modifier de manière plus ou moins importante. La figure 1.1 représente une vue globale de l'appareil phonatoire à gauche et à droite une section du larynx.

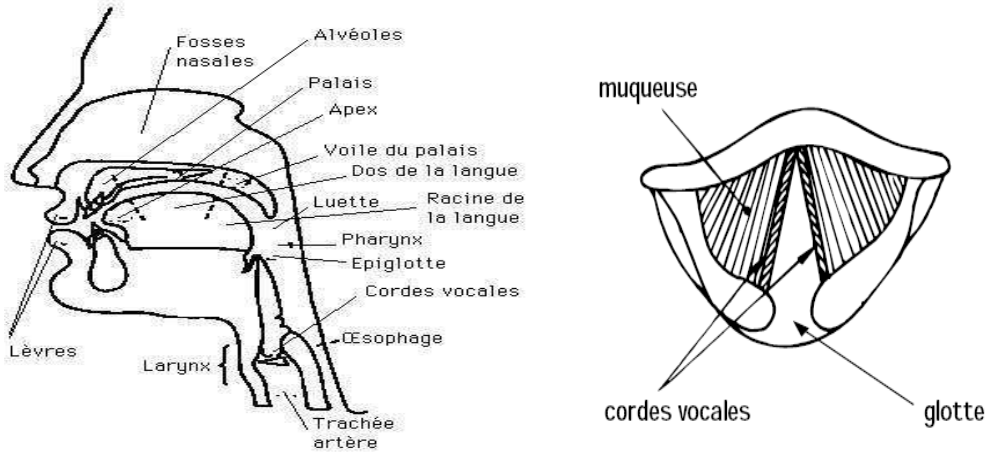


Figure 1.1 – Le système vocal humain

Le larynx se compose de 4 cartilages différents, dont le cartilage thyroïdien (pomme d'Adam) et l'épiglotte (cartilage en forme de lame, pouvant fermer par un mouvement de bascule en arrière l'entrée du larynx afin d'empêcher le bol alimentaire d'entrer dans le larynx et la trachée-artère). À l'intérieur du larynx se situent les *cordes vocales*, organes vibratoires constitués de tissu musculaire et de tissu conjonctif résistant. Les cordes vocales sont reliées à l'avant au cartilage thyroïdien. Elles peuvent s'écarter ou s'accoler pour produire des ondes de pression. L'espace entre les cordes vocales est appelé *glotte*. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés (ou sourds[1]). Les sons voisés (ou sonores) résultent au contraire d'une vibration périodique des cordes vocales.

L'air laryngé passe dans le conduit vocal qui mesure en moyenne entre 17 et 18 cm chez un sujet adulte et un peu moins pour une femme ou un enfant. Le conduit vocal

[1]Les phonéticiens appellent sourd ou sonore ce que les ingénieurs qualifient de voisé ou non voisé.

comprend plusieurs cavités supra-glottiques reliées entre elles : le pharynx, les fosses nasales, la bouche et les lèvres.

Le pharynx (cavité pharyngale) est un conduit musculo-membraneux situé entre la bouche et l'œsophage d'une part et entre les fosses nasales et le larynx d'autre part. La paroi du pharynx est constituée de muscles constricteurs. La constriction de ces muscles modifie le diamètre du pharynx. La racine de la langue peut également reculer ou avancer et donc agir sur le volume de cette première cavité supraglottique.

Les fosses nasales (cavité nasale) sont deux cavités cunéiformes séparées par une cloison verticale médiane et recouvertes de muqueuses. L'air passe par le nez lorsque le voile du palais (prolongement musculaire du palais osseux) est rabaissé (passage oro-nasal ouvert).

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

Les lèvres forment la cavité labiale lorsqu'elles sont projetées en avant (protrusion labiale).

1.2.2 Niveaux phonétique et phonologique

La majorité des sons du langage sont le fait du passage d'une colonne d'air venant des poumons, qui traverse plusieurs résonateurs de l'appareil phonatoire. Lors de l'acte parlé, la forme et la position de ces résonateurs varient de façon continue. A ce continuum observable dans l'espace des réalisations acoustiques correspondent des classes des sons. C'est ce que tente de faire ressortir la phonétique et la phonologie sous deux angles différents.

La phonétique s'intéresse aux sons eux-mêmes (production, transmission et perception), indépendamment de leur fonctionnement les uns avec les autres, tandis que la phonologie étudie les principes qui régissent l'apparition des sons et comment ils s'organisent afin de former les énoncés d'une langue donnée.

L'unité du codage linguistique de la phonologie est le phonème. Contrairement à un son, qu'on peut entendre et mesurer, un phonème est une entité abstraite, une classe de sons qui partagent la même opposition à d'autres sons dans une langue. Dans la

transcription, les phonèmes sont distingués par rapport aux sons par l'utilisation de barres obliques plutôt que des crochets. [b] est un son, mais /b/ est une classe de sons ou phonème.

La liste des phonèmes pour une langue donnée est établie sur la base de l'étude de *paires minimales*, composées de paires de mots différant par un seul son, lequel suffit à changer leur sens. Exemples : *zona* et *sauna* sont deux mots différents de la langue française, et il n'y a qu'un seul son différent (le premier). Les consonnes /s/ et /z/ sont donc deux phonèmes différents. En revanche, *roi* avec un /r/ roulé ([r]) et *roi* avec un /r/ non roulé ([R]) sont deux mots identifiés au même signifié [2]. Il n'y a donc pas d'opposition de sens entre le /r/ roulé et le /r/ non roulé, qui représentent alors le même phonème.

Les phonèmes apparaissent sous une multitude de formes articulatoires, appelées *allophones* (ou variantes). Ces derniers sont la réalisation acoustique (prononciation) d'un phonème selon l'environnement phonétique, qui conditionne (transforme) souvent la prononciation de ce son particulier. La liste des étiquettes de l'alphabet phonétique du système de synthèse de France Télécom R&D ainsi que les étiquettes correspondantes dans l'alphabet phonétique international sont présentées dans le tableau 1.1.

Les phonéticiens regroupent les sons de parole en deux grandes classes phonétiques en fonction de leur mode articulatoire : les voyelles et les consonnes.

La caractéristique majeure des voyelles est le libre passage de l'air à partir des cavités supraglottiques. Au cours de la propagation de l'onde glottique, des phénomènes de résonances vont entrer en jeu et modifier le contour spectral du signal glottique. On dénombre trois résonateurs (labial, buccal et nasal) qui, en fonction de leur forme, vont caractériser les voyelles prononcées. Contrairement aux voyelles, les consonnes sont produites lorsque le passage de l'air venant des poumons est partiellement ou totalement obstrué. Il existe deux grands types d'articulations consonantiques. La première apparaît quand le passage de l'air est fermé et le son résulte de son ouverture subite ; on a alors affaire à des *occlusives*. La seconde se produit quand le passage se rétrécit mais n'est pas interrompu ; on parle dans ce cas de continues, dont les *fricatives* sont les plus représentatives.

[2] Un signe linguistique est une entité formée par la réunion d'un signifié (un concept) et d'un signifiant (une forme sonore ou image acoustique). Par exemple, le mot français arbre est un signe linguistique associant le concept d'arbre à la forme sonore /arbr/.

Consonnes			
	Labiales	Dentales	Palatales
Plosives non voisées	P /p/ **p**ot	T /t/ **t**u	K /k/ **q**ui
Plosives voisées	B /b/ **b**eau	D /d/ **d**o	G /g/ **g**ai
Fricatives non voisées	F /f/ **f**a	S /s/ **s**a	CH /ʃ/ **ch**ez
Fricatives voisées	V /v/ **v**eau	Z /z/ **z**éro	J /ʒ/ **j**e
Nasales	M /m/ **m**on N /n/ **n**ez		
Liquide	L /l/ **l**e R /r/ **r**ien		
Semi-voyelles			
W /w/ l**oi**n Y /j/ b**ie**n			
Diphtongue			
UI /ɥi/ nu**i**t			
Voyelles			
I /i/ t**i**c	U /y/ l**u**	OU /u/ c**ou**p	IN /ɛ̃/ p**ain**
EI /e/ cl**é**	EU /ø/ p**eu**	AU /o/ p**o**t	UN /œ̃/ br**un**
AI /ɛ/ s**ei**ze	OE /œ/ l**eu**r	O /ɔ/ p**o**mme	AN /ã/ v**ent**
A /a/ l**à**	E /ə/ null**e**	ON /ɔ̃/ b**on**	
Silence			
# (sil, pau)			

Tableau 1.1 – Alphabet phonétique du Français de France Télécom R&D, les étiquettes A.P.I correspondantes sont entre /./

1.2.3 Niveau acoustique

La parole est le résultat d'une stimulation des cavités supraglottiques (conduit oral, conduit nasal) par un signal acoustique créé par le flux d'air en provenance des poumons et modulé par les cordes vocales. Les modèles les plus classiques de représentation du signal de parole (modèles de type source-filtre) s'inspirent de ce mode de production. Le signal de source résulte de la production d'une onde acoustique au niveau de la glotte. Cette onde passe ensuite dans le conduit vocal (oral, nasal) et subit l'effet de radiation des lèvres. Les transformations du signal de source par ces différents organes peuvent être modélisées par un simple filtrage linéaire. La figure 1.2 présente respectivement le signal de source, le filtre et le signal de parole dans les deux domaines fréquentiel et

temporel.

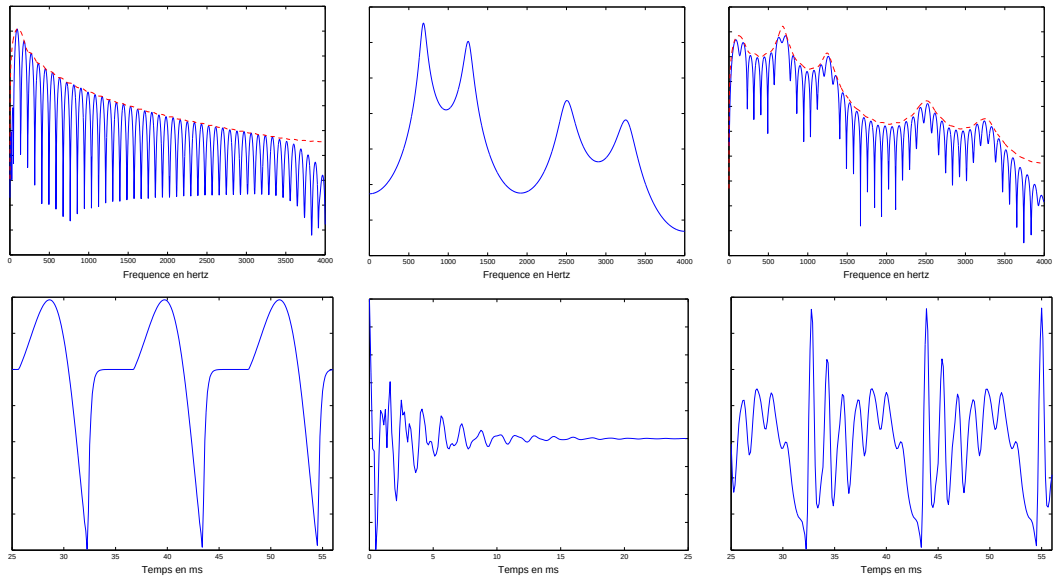


Figure 1.2 – Production de la parole : le modèle source-filtre présenté dans les domaines fréquentiel et temporel

Modélisation de la source glottique

La modélisation de la source dépend du type de son considéré (voisé ou non voisé), du mode de phonation (chuchoté, crié...) et de l'effort vocal. Un son voisé est un signal quasi périodique résultant du mouvement d'ouverture et de fermeture des cordes vocales modulant le débit d'air s'écoulant à travers la glotte. Le modèle de cette onde de débit glottique est contrôlé par quelques paramètres (période de vibration ou fréquence fondamentale[3], coefficient d'ouverture, coefficient de fermeture, amplitude...). Lorsque les cordes vocales n'entrent pas en vibration, un flux d'air passe librement. Ceci se traduit par la production de sons non voisés assimilables au niveau de la glotte à un bruit blanc. Pour ces sons les cordes vocales restent ouvertes pour permettre le passage d'un flux d'air en provenance des poumons ; l'excitation est due soit au relâchement rapide d'une occlusion complète du conduit vocal (plosives), soit aux turbulences du flux d'air créées au passage d'une constriction du conduit vocal (fricatives). Ces signaux sont modélisés par des sources de bruit réparties dans le conduit vocal, dont la position

[3]la fréquence fondamentale correspond à la fréquence de vibration des cordes vocales. Le pitch désigne quant à lui la hauteur perçue d'un son. Dans cette thèse nous ne ferons pas de distinction entre ces deux termes

et la puissance sont contrôlées.

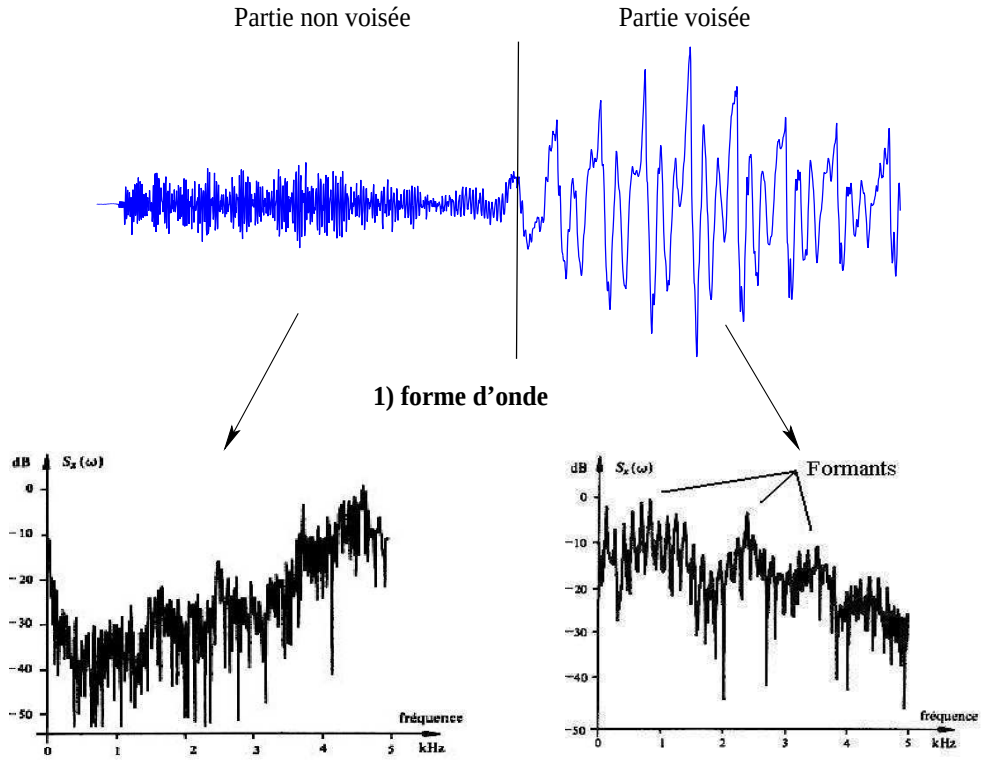


Figure 1.3 – Signal de parole

Les caractéristiques du filtre

Le conduit vocal est considéré comme un filtre ayant une fonction de transfert composée dans le domaine fréquentiel de plusieurs résonances dites formants. Ces derniers sont induits par les résonances propres aux différents volumes qui composent le conduit vocal. Cette dénomination de formant tient au fait que ce sont les résonateurs qui mettent en forme le signal glottique. Notons également que lorsque le passage oral-nasal est ouvert, la mise en parallèle du conduit nasal avec le conduit vocal se manifeste spectralement par l'apparition d'anti-résonances appelées aussi anti-formants.

Les caractéristiques acoustiques des cavités supra-glottiques peuvent être modélisées à l'aide d'un filtre linéaire AR (autorégressif) dont la fonction de transfert s'exprime comme suit :

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \tag{1.1}$$

où les a_i sont les coefficients de prédiction du filtre.

Pour une parole intelligible, le nombre de coefficients a_i est fixé de telle façon que la fonction de transfert du filtre présente un nombre suffisant de résonances pour modéliser correctement les 3 à 5 premiers formants des segments voisés. La figure 1.3 présente les parties voisée et non voisée d'un signal de parole ainsi que leurs densités spectrales de puissance respectives. La structure formantique de la partie voisée est également présentée sur la figure 1.3.

1.3 Les systèmes de synthèse de la parole

Les systèmes de synthèse de la parole peuvent être classés en deux familles correspondant à deux technologies majeures. La première est la synthèse *à partir de concepts*. Elle consiste en la formation en langage naturel de requêtes ou de connaissances que possède un système sur son environnement. La seconde approche, concernée par cette thèse, est la synthèse *à partir du texte*. Elle reçoit en entrée une forme textuelle suivant un format variable. Elle doit alors effectuer une analyse complexe de ce texte pour essayer d'en déterminer la structure linguistique sous-jacente puis produire le son. Cette approche peut être utilisée dans une large gamme d'applications, comme l'apprentissage des langues, des application de dialogue homme machine, la vocalisation de courrier électronique, etc...

1.3.1 Architecture d'un système de synthèse

Tout système de synthèse de parole à partir du texte (dit également TTS, de l'anglais " text-to-speech ") est généralement constitué de deux blocs de traitements principaux : un bloc de traitements linguistiques et un bloc de traitements acoustiques.

Le premier bloc vise à analyser et à structurer le texte afin de déterminer un mode de prononciation cohérent, puis à transformer le texte analysé en une séquence de descripteurs symboliques décrivant les unités cible. Le deuxième bloc consiste à générer un signal acoustique adapté à cette séquence symbolique.

La figure 1.4 représente une architecture classique d'un système de synthèse de la parole à partir du texte. Elle ce compose de deux blocs de traitements cités en introduction (traitement linguistique et traitement acoustique). Le premier bloc est composé de trois modules principaux qui permettent de transformer la forme textuelle du message à synthétiser en une chaîne symbolique, en général les phonèmes [4], munie d'indications prosodiques caractérisant l'élocution (durée des différents sons et des pauses, évolution de la mélodie). Cette représentation phonético-prosodique est ensuite utilisée par l'étage de synthèse sonore, qui assure la génération du signal de parole. Bien que tout ce qui sera cité concernera le français, il est important de souligner que cette architecture est applicable pour tout type de langues. La suite de ce chapitre va approfondir les objectifs de ces deux étapes en mettant l'accent sur les parties concernées par cette thèse.

1.3.2 Méthodes de génération du signal de parole

Deux groupes d'approches de génération du signal de parole coexistent jusqu'à présent : celles qui cherchent à modéliser le signal de parole en se basant sur une modélisation de l'appareil phonatoire figure 1.1 et celles qui visent à produire un signal de parole par concaténation de segments pré-enregistrés. Parmi la première classe se trouvent les approches de génération à partir de règles et les techniques de synthèse articulatoire, qui génèrent le signal de parole uniquement à partir d'informations paramétriques.

1.3.2.1 Synthèse par règles

Cette approche est fondée sur un modèle de production du signal vocal, modèle commandé par un nombre restreint de paramètres. La synthèse se décompose alors en deux étapes : une transformation des informations phonético-prosodiques, à l'aide

[4] En théorie de la production du signal de la parole, le terme "phonème" désigne l'unité acoustique minimale.

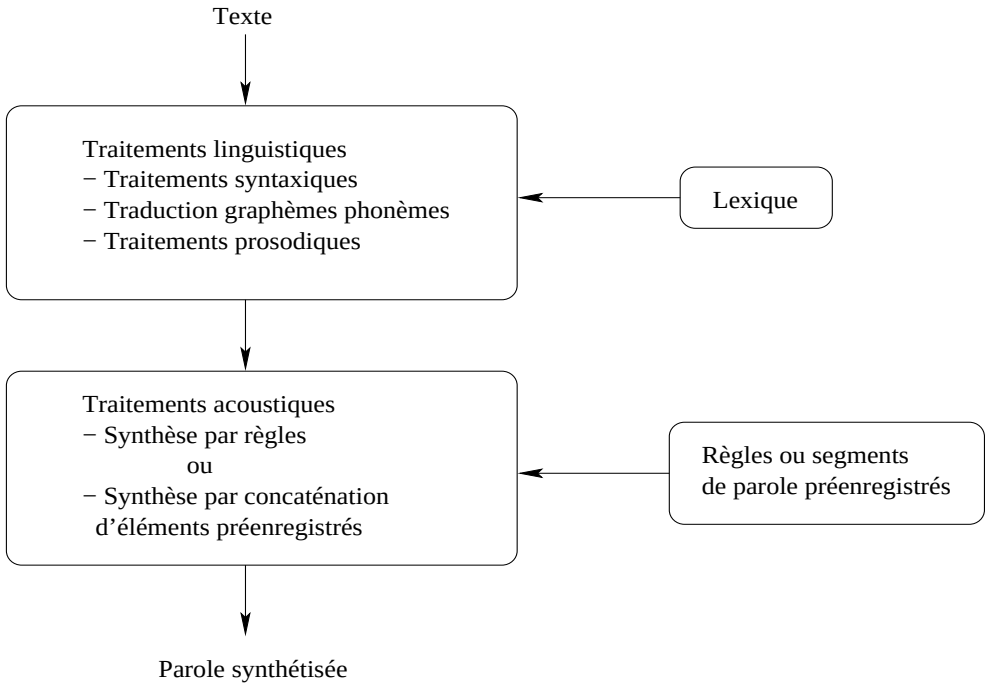


Figure 1.4 – Architecture d'un système de synthèse de la parole à partir du texte

de règles contextuelles [Kla79], en commandes permettant de spécifier l'évolution temporelle des paramètres du modèle de synthèse ; les paramètres ainsi déterminés sont utilisés pour synthétiser le signal acoustique.

Dans ce type de synthèse, les caractéristiques supra-glottiques sont modélisées à l'aide d'un filtre linéaire dont la fonction de transfert varie au cours du temps. Les paramètres utilisés pour le contrôle du filtre sont les paramètres formantiques, à savoir la fréquence centrale, la bande passante et l'amplitude des maxima significatifs de la fonction de transfert du conduit vocal comme présenté à la figure 1.3. Pour obtenir une parole intelligible, il suffit de spécifier les paramètres des 3 à 4 formants les plus importants, d'où la dénomination de synthèse par formants couramment employée pour ce type de synthèse. Une telle approche ne permet pas de restituer un signal de parole apparaissant naturel. La qualité médiocre obtenue résulte d'une part de la difficulté à modéliser suffisamment finement les trajectoires acoustique et d'autre part de la modélisation trop grossière du signal glottique.

1.3.2.2 Synthèse articulatoire

Cette technique est basée sur une modélisation géométrique du conduit vocal. Elle consiste à représenter le conduit vocal comme un tube de section variable, avec des embranchements et des sections parallèles, puis à y simuler le trajet des ondes produites au niveau de la glotte. Les modèles d'écoulement d'air (mécanique des fluides), de sources et de propagation acoustique (phénomènes physiques), en association avec des modèles articulatoires (mécaniques), permettent de constituer un synthétiseur articulatoire complet, contrôlé par deux jeux de paramètres : les paramètres supra-laryngés qui commandent le modèle articulatoire, et un jeu de paramètres qui pilotent les cordes vocales (pression sub-glottique, longueur des cordes vocales et hauteur de la glotte au repos) [Mae79].

Ces deux techniques de génération du signal de la parole (synthèse articulatoire et par règles) n'étant pas directement liées à notre thèse, nous ne détaillerons pas ces approches dans ce manuscrit. Pour de plus amples informations sur les travaux en synthèse articulatoire et par règles, le lecteur pourra se référer à [Gab94] et [SK94].

1.3.2.3 Synthèse par concaténation d'unités acoustiques

Contrairement aux deux approches citées précédemment, la synthèse par concaténation d'unités acoustiques ne fait pas explicitement, tout au moins dans son principe, référence à un modèle de production de la parole. Elle fonctionne en concaténant des unités acoustiques, c'est-à-dire en mettant bout à bout des signaux de parole pré-enregistrés. Cette technique est la seule qui permette à ce jour de synthétiser de la parole dont le timbre est proche de celui d'un locuteur humain. La suite de ce document traite uniquement de la synthèse par concaténation et nous détaillons dans les section 1.4 et 1.5 respectivement les traitements linguistiques et acoustiques mis en œuvre dans un tel système.

1.4 Traitements linguistiques

Le bloc de traitements linguistiques regroupe les différents modules qui permettent de transformer la forme textuelle du message à synthétiser en une chaîne de phonèmes éventuellement enrichis d'informations linguistiques et prosodiques caractérisant l'élocution.

Ces différents modules sont : les prétraitements des éléments non lexicaux, l'analyse lexicale, l'analyse syntaxique, la transcription orthographique-phonémique et le traitement prosodique.

1.4.1 Prétraitement des éléments non lexicaux

Cette étape de prétraitement a pour objet de retranscrire en toutes lettres les chaînes non orthographiques, c'est-à-dire celles qui ne sont pas uniquement constituées de caractères orthographiques. Il peut s'agir de chiffres, de dates (29/08/99, 29 Jan. 1999) ou plus généralement de sigles composés de caractères orthographiques et numériques (vol AF1024, référence SD44). En général, des règles de transcription sont utilisées pour le traitement des quantités numériques, des dates ou des sigles standards (SNCF, PTT, etc ...). Si le système de synthèse est destiné à un domaine spécifique, le lexique propre à ce domaine sera appliqué.

1.4.2 Analyse lexicale

L'analyse lexicale consiste à déterminer dans un lexique les différents lexèmes[5] composant le texte orthographique à synthétiser. Cette analyse est réalisée en trois étapes : un découpage du texte en lexèmes, une analyse morphologique et une analyse lexicale.

Lors du découpage du texte en lexèmes, chaque lexème se voit attribuer une ou plusieurs catégories grammaticales potentielles (nom, adjectif, verbe, adverbe, pronom, etc...), éventuellement augmentées d'informations relatives aux propriétés grammaticales (genre, nombre, conjugaison, verbe d'état, etc...).

Exemple : Un mot aussi anodin que "voile" peut être simultanément un nom masculin (morceau d'étoffe destiné à cacher une ouverture, un monument, un visage... : "porter le voile"), un nom féminin (morceau de forte toile qui reçoit l'action du vent : "larguer les voiles"), un verbe transitif à l'indicatif présent ou au subjonctif présent (action de cacher d'un voile au sens propre ou au sens figuré : "voiler la vérité"), un verbe pronominal (perdre de son éclat, se ternir : "le ciel se voile" ; se dit aussi d'une "roue qui se tord légèrement").

[5]Dans ce contexte le terme "lexème", qui représente une suite de caractères orthographique, est plus approprié que "mot".

L'analyse morphologique a pour objet de décomposer le lexème en composantes élémentaires, les morphèmes, correspondant aux préfixes, suffixes, désinences (marques du féminin ou du pluriel pour les noms et les adjectifs, temps, personne et mode pour les verbes), racines. La racine de chaque lexème est repérée (par exemple "affreux" pour "affreusement") et stockée dans le dictionnaire à la place de tous ses dérivés. Certaines séquences de lexèmes peuvent également être intégrées au sein de syntagmes[6] courts lexicaux (par exemple "salle de bains"), de telles séquences étant plus faciles à traiter comme de simples unités que comme plusieurs lexèmes séparés. En résumé, cette analyse a trois fonctions principales :

- former certaines unités lexicales usuelles ;
- retrouver la racine de chaque lexème ;
- déterminer la catégorie grammaticale de chaque lexème.

L'analyse lexicale effectue une première étape de phonétisation en associant à chaque lexème présent dans le lexique sa transcription phonétique. À cet effet, un véritable "alphabet phonétique", issu de l'alphabet phonétique international, est utilisé 1.1.

1.4.3 Analyse syntaxique

L'analyse syntaxique vise à déterminer la structure de la phrase. Elle est conduite par application de règles, ces règles pouvant être de deux types. Dans certains cas, il peut s'agir d'heuristiques, résultant généralement de l'application de règles grammaticales standards (par exemple, on ne peut observer la succession de deux verbes conjugués). En complément ou à la place de ces heuristiques parfois très complexes, on utilise aussi fréquemment des règles probabilistes, exploitant des modèles de langage. Ces modèles sont fondés sur l'observation que toutes les séquences de catégories grammaticales dans une langue donnée ne sont pas équiprobables. On peut donc tenter de résoudre les ambiguïtés en recherchant dans l'ensemble des séquences possibles de catégories grammaticales (chaque mot est *a priori* porteur de plusieurs catégories possibles et l'on considère l'ensemble des transitions entre ces différentes catégories) la séquence de catégories la plus probable.

La connaissance de la catégorie syntaxique exacte est également utile pour déterminer la prononciation correcte et notamment pour désambiguïser les homographes hétérophones.

[6] un ensemble de mots qui sont tous en relation avec un élément central appelé "noyau".

Considérons par exemple la phrase : "Les poules du couvent couvent".
Dans le premier cas, *couvent* est un nom commun prononcé /kuvã/, tandis que dans le second, c'est le verbe *couver* à l'indicatif présent et se prononce /kuv/. La première tâche de l'analyse est donc d'assigner une et une seule catégorie grammaticale à chaque lexème à partir des résultats fournis par l'analyse lexicale précédente.

1.4.4 Transcription orthographique-phonétique

Traditionnellement appelée *conversion graphème-phonème*, l'étape de transcription orthographique-phonétique constitue le noyau minimal, indispensable à tout système de synthèse de parole, aussi élémentaire soit-il. Cette étape repose sur l'utilisation d'un automate paramétré appliquant un ensemble de règles de réécriture, qui permettent d'associer un phonème (ou un groupe de phonèmes) à un caractère (ou un groupe de caractères) orthographique en prenant en compte le contexte gauche et le contexte droit. Ces règles sont organisées de façon hiérarchique, des règles les plus particulières aux règles les plus générales.

Le nombre de règles nécessaires pour effectuer la transcription orthographique-phonétique dépend de la langue que l'on considère ; si on prend le cas de la langue espagnole, où la forme orthographique est très proche de la forme phonétique, le nombre de règles requises est de moins de 100 règles. Par contre, un système minimal de description des règles de phonétisation du français standard se compose environ de 500 règles.

Exemple : le mot "oiseau" se transcrit phonétiquement /wazo/, par application des règles suivantes.

1. La chaîne de caractères orthographiques "oi" se transcrit par la succession des phonèmes /wa/, parce qu'elle est précédée d'un séparateur de mot et qu'elle n'est pas suivie de la chaîne "gn" comme dans "oignon", ou d'un "n" comme dans "oindre".

2. La lettre "s" se transcrit par le phonème /z/ car cette lettre est entourée par deux voyelles et que "oiseau" ne fait pas partie d'une liste d'exceptions à cette règle, stockée dans le lexique (on pense en particulier à "paraSol" ou " vraiSemblance").

3. La chaîne de caractères "eau" se transcrit par le phonème /o/, indépendamment du contexte.

1.4.5 Traitements prosodiques

La prosodie est l'étude des phénomènes de l'accentuation, de l'intonation et du rythme (variation de hauteur, de durée et d'intensité) permettant de véhiculer de l'information liée au sens telle que la mise en relief, mais aussi l'assertion, l'interrogation, l'injonction ou l'exclamation.

En l'état actuel du savoir faire, les paramètres prosodiques prédits sont le rythme (ensemble des durées des segments phonétiques et des pauses), l'intonation (type de voisement, et pour les segments voisés, les valeurs de fréquence fondamentale) et l'énergie du signal.

Tout comme la phrase s'ordonne de façon hiérarchique en groupes syntaxiques ou syntagmes[7] (le syntagme sujet, le syntagme verbal, le syntagme complément), la phrase prosodique s'organise en une hiérarchie complexe de groupes prosodiques, groupes de mots le plus souvent séparés par des pauses. Sur chacun de ces groupes prosodiques, les paramètres prosodiques suivent une évolution particulière, dépendant du rôle du groupe prosodique dans la phrase, de ses dépendances fonctionnelles avec les groupes adjacents, du nombre de syllabes, mais aussi du sens de la phrase et de l'intention du locuteur. Les frontières des groupes prosodiques ne coïncident pas systématiquement avec les limites de groupes syntaxiques. Une certaine congruence peut toutefois être notée, surtout pour ce qui concerne les frontières syntaxiques majeures (frontière de phrase ou de clause, mais aussi frontière entre le syntagme sujet et le syntagme verbal associé).

Par exemple, les expressions telles que "le père missionnaire" et "le permissionnaire" sont, du point de vue de la chaîne sonore, tout à fait identiques. La mise en relief de certaines syllabes (accentuation), les modulations de la hauteur de la voix (mélodie) et la présence de pauses sont autant d'indices qui permettront à l'auditeur de savoir exactement quelle interprétation il doit donner à cette suite de sons. La chaîne parlée, qui est constituée d'une succession linéaire de segments vocaliques et consonantiques, est d'abord subdivisée en unités suprasegmentales qui facilitent le décodage du message par l'auditeur. La délimitation de ces unités est faite à l'aide de marqueurs dont la réalisation fait appel à des variations paramétriques de durée, de fréquence et d'in-

[7] Egalement appelé groupe, un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle.

tensité. La qualité de ces marqueurs prosodiques est étroitement liée à la qualité des différents traitements linguistiques, qui permettront d'effectuer, à l'aide d'heuristiques et de règles, le passage d'une information symbolique vers une information prosodique numérique.

Les traitements prosodiques sont complexes et s'articulent en différents modules (insertion des pauses, durées phonétiques et fréquence fondamentale), décrits brièvement ci-après. En revanche, l'avènement des approches de synthèse par sélection dynamique d'unités non uniformes de segments de parole permettent d'envisager des techniques nouvelles pour la génération de la prosodie. En effet, ces approches génèrent automatiquement la prosodie sans modèle a priori puisqu'elles utilisent une caractérisation symbolique fine des unités d'un corpus de grande taille, ce qui permet de conserver la prosodie originale des segments sélectionnés.

1.4.5.1 Insertion des pauses

Les pauses correspondent aux silences, de durées variables, qui s'insèrent à la fin de chacun des groupes de souffle. L'importance de la coupure syntaxique liée à un marqueur syntaxico-prosodique détermine la durée de la pause à insérer. Ce facteur est particulièrement important pour le naturel de l'élocution.

1.4.5.2 Durées phonétiques

Une bonne détermination des durées est cruciale pour assurer le naturel de l'élocution. Des durées erronées produisent une parole heurtée, chaotique et parfois difficilement intelligible.

Deux approches existent, pour la modélisation de la durée. La première basée sur des règles et une bonne analyse statistique, initiée par [Kla79], détermine la durée en prenant en compte différents facteurs, en particulier la durée intrinsèque des sons constituant le segment et le contexte. Parmi les facteurs influençant la durée phonétique, nous pouvons citer : le contexte phonétique (certains phonèmes ont tendance à allonger les phonèmes adjacents, d'autres auront tendance à les raccourcir), la position de la syllabe porteuse dans le groupe prosodique (en français par exemple, la syllabe finale des mots est généralement allongée, d'un facteur d'autant plus important que le groupe

précède une frontière syntaxique majeure), la nature du groupe prosodique (sa fonction dans la phrase), la longueur du groupe prosodique, etc...
Ces règles sont souvent déterminées de manière experte, et requiert une longue phase d'analyse de corpus de parole. De plus, l'ensemble de règles ainsi déterminé n'est valide que pour un locuteur donné.

La deuxième approche est basée sur des techniques d'apprentissage automatique. Celles-ci peuvent reposer sur l'utilisation de réseaux connexionnistes pour prédire la durée des syllabes et ainsi calculer les durées des phonèmes à partir de leur moyenne et de leur écart-type comme dans [CI91] pour l'anglais ou [Tou98] pour le français. Dans [PWOB90], Price et al. proposent un modèle HMM[8] à 7 états pour détecter automatiquement les coupures prosodiques à partir de l'analyse des durées des phonèmes.

1.4.5.3 Fréquence fondamentale

Le contrôle de la fréquence fondamentale, dont l'évolution dans le temps définit le contour mélodique, est le point essentiel pour la détermination de l'intonation. L'évolution de la fréquence fondamentale pour chaque phonème est spécifiée à l'aide d'un modèle prédictif complexe, prenant en compte deux types de phénomène : des phénomènes locaux dits *de micromélodie*, des phénomènes globaux dits *de macromélodie*.

La micromélodie est l'influence de l'évolution de la fréquence fondamentale par sa position dans la syllabe et par son environnement phonétique immédiat (certains phonèmes, comme les consonnes occlusives, contribuent à abaisser la fréquence fondamentale ; d'autres ont tendance à l'augmenter). La macromélodie a une portée supérieure à celle de la syllabe (groupe prosodique, phrase). Les facteurs influant sur la macromélodie sont la position de la syllabe dans le groupe prosodique, la fonction du groupe prosodique dans la phrase et le mode de la phrase (interrogatif, déclaratif...).

Plusieurs modèles de contours mélodiques ont été proposés. Dans [FH82], Fujisaki et Hirose utilisent un modèle acoustique source/filtre d'un ensemble limité de paramètres structurés entre eux pour modéliser la fréquence fondamentale. Taylor dans [Tay93] propose une modélisation automatique de la courbe de fréquence fondamentale en terme de montées et de descentes non linéaires et de connexions.

[8] modèle de markov caché, "Hidden Markov Model" en anglais

1.5 Traitements acoustiques

Les traitements acoustiques prennent en entrée un ensemble d'informations résultant des différents traitements linguistiques (séquence phonétique annotée sur les plans linguistique et prosodique). Ces traitements impliquent une correspondance entre une représentation symbolique abstraite (unité symbolique) et une réalisation acoustique (unité acoustique). Cette correspondance représente le point de démarquage entre le processus impliqué dans les traitements linguistiques et celui chargé des traitements acoustiques. Dans la suite de ce document, quand le mot *unité* est cité tout seul, il désigne l'unité symbolique. La figure 1.5 montre les trois principaux modules de génération sonore d'un système de synthèse par concaténation d'unités acoustiques.

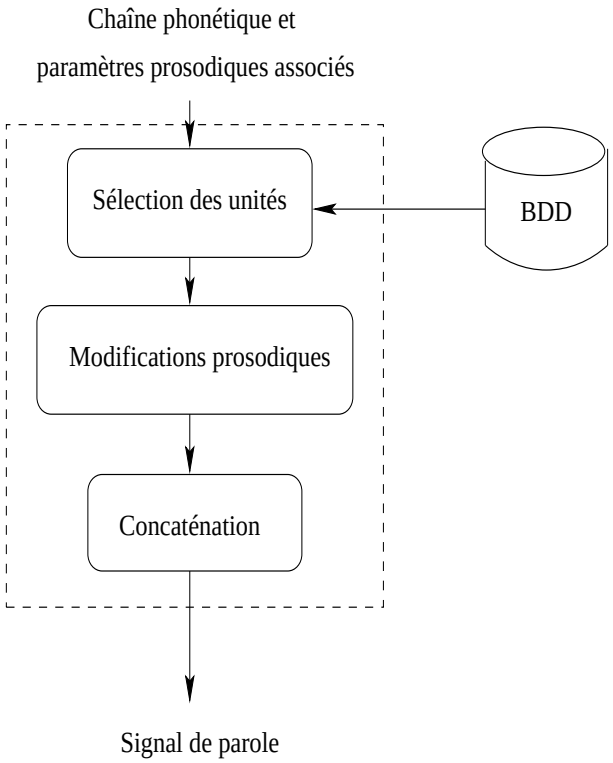


Figure 1.5 – Architecture générale des traitements acoustiques dans un système de synthèse par concaténation

Dans la suite de cette section, nous décrivons les différentes unités symboliques utilisées dans un système de synthèse par concaténation et nous présentons les modules de traitements acoustiques.

1.5.1 Types d'unités de synthèse

Dans le cadre de la synthèse par concaténation d'unités acoustiques, le choix des unités joue un rôle primordial. La variation des unités peut se faire sur deux plans : en longueur (phonème, diphonème, etc...) et au niveau de ses réalisations acoustiques (mono-représentées, multi-représentées). Une unité multi-représentée signifie que cette même unité est présente plusieurs fois dans le corpus, chaque instance de l'unité se distinguant des autres au niveau acoustique. Ces différence acoustiques peuvent avoir un impact sur le plan de la perception. Les principaux types d'unités acoustiques (phone, diphone, etc...) sont représentés dans la figure 1.6.

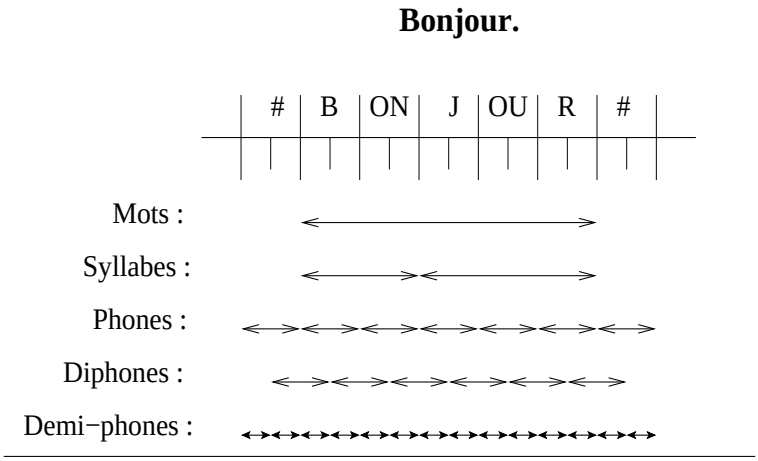


Figure 1.6 – Les principaux types d'unités

1.5.1.1 Les phones et leurs dérivés

Certaines unités acoustiques, par exemple les phones[9], comme l'avait montré [Har53], sont inappropriés car ils ne permettent pas de capturer la dynamique du processus de production de parole. Contrairement à ce que pourrait laisser croire la théorie linguistique, la parole est essentiellement un processus continu et l'enchaînement des sons entre eux (continuité inter-unités qui n'est rien d'autre que la manifestation acoustique de l'articulation) est au moins aussi important, du point de vue de la perception, que les sons eux-mêmes (continuité intra-unités). Considérons par exemple les deux mots < épiler > et < épauler >, qui peuvent respectivement être transcrits phonétiquement

[9] réalisation acoustique d'un phonème.

en /epile/ et /epole/. Un certain nombre de phénomènes de co-articulation y sont à l'œuvre. Les voyelles /i/ et /o/ sont affectées par le son /p/ les précédant, ce qui les rend légèrement aspirées à l'initiale, et elles sont également labialisées de par le son /l/ les suivant. De tels effets de co-articulation soulignent le besoin d'enregistrer des allophones, des phones dans leurs contextes gauche et droit [MO86]. Ainsi, par une approche basée sur une distance acoustique, Nakajima dans [NH88] sélectionne dans sa base de donnée les allophones les plus typiques de classes contextuelles.

Certains auteurs comme Conkie, dans l'objectif d'accroître le contrôle sur la nature des concaténations des phones, utilisent des unités acoustiques dérivées du phone, *les demi-phones* [Con99]. Ces derniers permettent de procéder à des concaténations sur des frontières de phones dans les zones de faible coarticulation et privilégie les concaténations sur des zones stables lorsque ces effets de coarticulations sont trop importants.

En outre, s'inspirant des formalismes de la reconnaissance automatique de la parole, Donovan dans [Don96] et Huang dans [HAA+96] utilisent des unités acoustiques relatives à un état de modèle de Markov caché, dites *sénones*.

1.5.1.2 Les diphones et leurs dérivés

L'unité minimale permettant d'obtenir une synthèse de qualité acceptable est le diphone, qui est défini comme "le segment de parole qui s'étend de la zone stable d'une réalisation phonétique à la zone stable de la réalisation suivante et qui protège en son centre toute la zone de transition" [Eme77]. Le diphone, à l'inverse du phone, capture la transition entre les différentes cibles articulatoires associées aux phones, transitions qui sont cruciales pour la perception des différents sons. En théorie, le nombre de diphones est égal au carré du nombre de phonèmes, soit 36^2 pour la langue française. En tenant compte du fait que certaines transitions entre phonèmes sont impossibles en français, le nombre effectif de diphones pour le français est de 1300. En pratique, pour la synthèse par diphones, le nombre d'unités utilisées est légèrement plus important (de l'ordre de 1 500 à 2 000) pour tenir compte des différentes variantes contextuelles des phonèmes composant le diphone. Ainsi, dans les dictionnaire minimal de diphones on est amené à rajouter des unités de type polyphone (regroupant au minimum 2 diphones) de façon à protéger de la concaténation certains phonèmes sensibles. Le volume de stockage

nécessaire est de l'ordre de 5 à 10 Mo (2 à 6 min de parole numérisée avec une fréquence d'échantillonnage de 16 kHz). Cette quantité de données (considérée dans les années 1990 comme une borne supérieure) semble aujourd'hui bien raisonnable au regard des possibilités de stockage offertes par les systèmes informatiques actuels.

La constitution du dictionnaire d'unités acoustiques se fait en enregistrant un corpus de logatomes, successions élémentaires de sons de parole dépourvue de signification. Avant la séance d'enregistrement les unités de base et leur contexte phonétique d'enregistrement sont spécifiés par un expert. Par exemple, le logatome AU_F_EU_T_EU permet uniquement d'extraire le diphone F_EU. L'extraction des unités acoustiques nécessite la segmentation des logatomes, tâche effectuée de manière semi-automatique. Dans un premier temps, un algorithme de segmentation automatique est appliqué, ce dernier étant basé sur des méthodes statistiques dérivées de la reconnaissance de parole (en général des HMM). Une étape de vérification manuelle est ensuite nécessaire pour corriger les frontières ainsi obtenues.

Dans un but d'isoler des phonèmes sensibles et fortement influencés par leur contexte phonétique, comme les liquides et les semi-voyelles, Salza et Sandri dans [SSF87] utilisent des unités acoustiques de type 2-diphones. Ces derniers améliorent la continuité inter-unités, en offrant des extrémités plus facilement concaténables que le diphone.

1.5.1.3 Multi-représentation des unités

Pour améliorer le naturel de la voix de synthèse, plusieurs auteurs se sont focalisés sur la variation en terme de réalisation acoustique et perceptive des unités. Ainsi, Sagisaka dans [Sag88] et Black dans [BC95] proposent de stocker, dans le but d'augmenter la continuité inter-unités de la séquence de synthèse, plusieurs allophones de même identité phonétique. Tout comme pour les allophones, il est possible d'améliorer la continuité inter-unités en stockant plusieurs diphones de même identité phonétique (choisis manuellement [SSF87], [TKTS02]).

L'avènement de ces approches a ouvert de nouvelles voies pour la génération de la prosodie. En effet, ces approches génèrent automatiquement la prosodie sans modèle a priori mais en se basant sur une caractérisation symbolique fine des unités acoustiques ce qui permet de conserver la prosodie originale des segments sélectionnés.

1.5.2 Architecture d'un système de concaténation

Comme montré sur la figure 1.5, les traitements acoustiques d'un système de synthèse par concaténation comprennent trois étapes distinctes présentées dans ce qui suit.

1.5.2.1 Sélection des unités acoustiques

Cette première étape consiste à choisir dans le répertoire d'unités acoustiques celles qui seront effectivement utilisées pour synthétiser la succession de sons désirée. Cette étape est relativement simple quand les unités sont mono-représentées (à l'instar des phonèmes et des diphones) : seule la présence de plusieurs versions pour le même segment est à prendre en considération. Cette étape est en revanche plus délicate pour les systèmes de synthèse par corpus, dont les unités sont multi-représentées. Pour une séquence d'unités symboliques donnée, plusieurs choix d'unités acoustiques sont en général possibles. Il faut alors arbitrer entre les différentes combinaisons d'unités acoustiques possibles et donc mettre en place un processus de sélection. Cette étape sera davantage détaillée dans le chapitre qui suit.

1.5.2.2 Ajustement des paramètres prosodiques

Les unités acoustiques préenregistrées possèdent une prosodie intrinsèque (les sons qui la composent ont une certaine durée et la fréquence fondamentale décrit un certain contour). Cette prosodie intrinsèque est exploitée dans les systèmes à sélection d'unités mais pour les systèmes à unités mono-représentées, la prosodie de synthèse sera spécifiée par le module prosodique. Dans ce cas, il est nécessaire d'utiliser une technique de traitement de signal pour ajuster aux valeurs cibles définies les paramètres prosodiques des unités de synthèse. Parmi ces techniques de modification prosodique étudiées jusqu'à présent, les plus efficaces peuvent toutes être considérées comme des variantes d'une même méthode : l'algorithme PSOLA[10] (pour Pitch-Synchronous Overlap-Add)[MC90].

1.5.2.3 Concaténation des unités

Dans un système par concaténation, les unités de parole sont sélectionnées afin d'être reliées pour former une séquence complète. Ces unités acoustiques, quelles que soient

[10]TD-PSOLA (Time-Domain PSOLA) est une marque déposée par France Télécom.

les précautions prises lors de leur sélection et de leur enregistrement, ne possèdent pas exactement à leur frontière les mêmes caractéristiques acoustiques. En l'absence de traitement, ces discontinuités vont engendrer des artefacts perceptibles et gênants. Pour les limiter, la solution classiquement employée consiste d'une part à ajuster les frontières de façon à minimiser les distances spectrales et d'autre part à opérer un lissage des segments à concaténer. Un tel lissage peut être obtenu par addition-recouvrement [HMC89], ou encore, lorsqu'une description paramétrique du signal de parole est disponible [LSM93], par un mécanisme d'interpolation de tout ou partie de ces paramètres acoustiques. Le lissage demeure néanmoins un traitement délicat en ce sens qu'il est difficile de pouvoir garantir un résultat satisfaisant. Par exemple, les techniques de lissage actuelle ne sont que moyennement efficaces pour corriger des défauts apparaissant lors de la concaténation de plosives voisées.

1.6 Conclusion

Ce chapitre a permis d'introduire certains concepts de base du traitement de la parole via une caractérisation du signal de parole sur les plans physiologique, acoustique et phonétique.

Le principales méthodes de synthèse de la parole ont été présentées et une attention particulière a été portée sur les approches par concaténation. Ainsi, nous avons passé en revue les différents types d'unités de synthèse utilisables et détaillé les différents modules de traitements linguistiques et acoustiques présents dans un système TTS par concaténation.

Dans cette présentation relativement générale nous avons introduit le principe de la synthèse par corpus qui fait l'objet de cette thèse. Nous proposons au chapitre 2 une étude plus détaillée de cette technique avant d'aborder aux chapitres suivants les principales contributions de cette thèse.

Chapitre 2

Synthèse par corpus (SPC)

2.1 Introduction

Au chapitre précédent, nous avons rapidement décrit la synthèse par corpus (SPC) comme étant une approche reposant sur l'utilisation de grandes bases de données de parole obtenues lors de l'enregistrement d'un locuteur professionnel. Notons d'emblée qu'avec l'avènement de cette technologie, la problématique de la synthèse par concaténation change radicalement. En effet, disposant d'un ensemble d'unités acoustiques plus riche qu'un simple jeu de diphones mono-représentés, il ne s'agit plus de satisfaire une cible prosodique en modifiant des segments de parole sans trop dégrader le timbre originel de la voix, mais plutôt de sélectionner les unités les mieux adaptées au contexte de synthèse et de les concaténer en faisant le minimum de traitements acoustiques. Cette stratégie peut se résumer par l'expression "choose the best to modify the least" [BPQ+99], c'est-à-dire qu'en choisissant les segments les mieux adaptés au contexte de synthèse, les modifications des unités seront limitées et, par conséquent, le timbre originel de la voix sera mieux préservé. De plus, moyennant une modélisation prosodique adéquate, la prosodie du locuteur peut également être convenablement restituée. Ainsi, la SPC permet d'obtenir une parole synthétique de très haute qualité et naturelle, dans le sens où elle respecte la variabilité intrinsèque du locuteur. À ce titre, elle s'est peu à peu imposée comme un standard dans le monde de la synthèse vocale. Les performances d'un système de SPC dépendent essentiellement de deux facteurs, à savoir d'une part la richesse du corpus de synthèse et d'autre part l'algorithmie de sélection mise en

place. Dans ce chapitre, nous commençons par décrire ce premier point en détaillant le processus de création de voix, puis nous dressons un état de l'art sur la sélection des unités en détaillant les types de critères de sélection classiquement employés ainsi que les méthodes de sélection existantes.

2.2 Processus de création de voix pour la SPC

Dans cette section, nous décrivons les différentes étapes nécessaires à la création d'une base de données directement utilisable dans un système de SPC.

2.2.1 Définition du corpus à enregistrer

Le choix du corpus est un élément clé pour la qualité d'un système de SPC. En effet, pour pouvoir espérer restituer la variabilité intrinsèque d'un locuteur, il convient au préalable de couvrir son univers de production. Par univers de production nous entendons l'espace acoustique et prosodique d'un locuteur. Cette espace est très difficile à qualifier, car les réalisations prosodiques et acoustiques d'un locuteur sont dans une large mesure dépendantes de son style d'élocution. Il est clair, par exemple, que la parole lue a une prosodie très différente de la parole spontanée dont le contenu expressif, voire émotionnel est autrement plus varié. Idéalement, un corpus de synthèse devrait être capable de couvrir l'étendue des styles d'élocution qu'un locuteur est susceptible de produire. Néanmoins, actuellement, les phénomènes paralinguistiques liés à la parole spontanée sont très difficiles à modéliser et donc à prendre en compte pour définir un corpus de synthèse. C'est pourquoi, la SPC se cantonne dans une large mesure à restituer un style de parole lue, porté essentiellement par la structure linguistique du texte à vocaliser. Étant donné ce cadre de travail, définir le corpus revient à déterminer l'ensemble des unités à enregistrer de façon à paver au mieux un certain espace acoustico-prosodique. Notons d'emblée que la résolution de ce problème sous-entend qu'un mécanisme d'inférence existe pour prédire les caractéristiques acoustiques et prosodiques d'une unité en fonction uniquement du texte. Pour cela, il est nécessaire de se donner un ensemble de descripteurs symboliques permettant de qualifier les unités sur les plans acoustiques et prosodiques. Un tel alphabet symbolique contient en général un ensemble assez riche d'informations linguistiques (type de syllabe, position de la syllabe dans le mot, etc...) voire prosodique (marqueurs mélodiques), de sorte qu'au final,

chaque unité peut être considérée selon de nombreuses configurations (au minimum 100 pour une description symbolique relativement basique). Ainsi, recenser l'ensemble de ces configurations en vue de les faire prononcer par un locuteur apparaît difficilement envisageable. Par exemple, dans le cas des diphones, cela supposerait d'enregistrer plusieurs centaines de milliers de mots ou d'expressions contenant ces unités. En outre, il s'agirait de bien vérifier que chaque unité a bel et bien été prononcée selon la configuration désirée et le cas échéant de procéder à un nouvel enregistrement des unités mal prononcées. Enfin, signalons également que le fait de contraindre fortement le locuteur à prononcer un ensemble d'unités relativement courtes (des mots voire des expressions courtes) selon des schémas prosodiques prédéfinis peut conduire à l'obtention d'une parole peu naturelle. Une façon plus réaliste et plus satisfaisante de procéder est de rechercher, parmi un vaste corpus textuel, un jeu minimal de phrases permettant une couverture symbolique acceptable. L'avantage de cette méthode est que la parole ainsi enregistrée aura été prononcée dans un cadre relativement peu contraint, d'où une meilleure préservation des caractéristiques intrinsèques du locuteur. Du point de vue algorithmique nous sommes alors face à un problème de recouvrement d'ensemble dont la complexité est de l'ordre de 2^N, où N désigne le nombre de phrases du corpus textuel d'origine. Pour résoudre un tel problème dit NP complet, des heuristiques doivent donc être mises en œuvre, comme dans [Fra02], où des approches par algorithmes gloutons ont été utilisées. Notons également qu'une difficulté particulière réside dans le fait que les fréquences d'apparition varient énormément d'une unité à l'autre, comme cela a été mentionné dans [vSB97]. Plus précisément, cela signifie que de nombreuses unités sont rares et que pour les recueillir un nombre trop important de phrases seraient nécessaires. Ces unités rares étant nombreuses, la probabilité de détecter une de ces unités lors de la synthèse d'une phrase quelconque est loin d'être négligeable. En résumé, malgré les efforts entrepris dans la constitution de corpus, la synthèse par corpus se heurte à des problèmes de "trou" de couverture. Bien qu'aucune étude sérieuse ne l'atteste réellement, il y a fort à parier que le manque de contrôle de la couverture de la base acoustique a des répercussions fâcheuses sur la qualité de la parole produite par un système de SPC.

2.2.2 Constitution d'un dictionnaire acoustique

Nous n'avons pour l'instant parlé que du corpus à enregistrer pour pouvoir créer une voix de SPC. Cette matière acoustique brute ayant été collectée, de nombreux traitements sont nécessaires avant d'aboutir à un dictionnaire acoustique exploitable par le système de synthèse. Nous présentons ci-après ces traitements tels qu'ils sont effectués à France Télécom. Tout d'abord, l'enregistrement brut doit être découpé en phrases. Ce processus peut dans une large mesure être automatisé puisque nous connaissons la séquence de phrases produites par le locuteur. La procédure de découpage consiste alors à fournir l'ensemble d'une session d'enregistrement, c'est-à-dire l'enregistrement sonore proprement dit ainsi que la séquence de textes prononcés, à un moteur de reconnaissance vocale. Compte tenu des performances des technologies de reconnaissance vocale actuelle cette méthode automatique donne en pratique de très bons résultats. Seules des corrections sur les frontières de phrases peuvent s'avérer nécessaires afin de mieux délimiter les silences de début et de fin de phrase. Une fois les phrases délimitées, la phonétisation est déterminée. Cette opération est semi automatique et consiste à corriger manuellement la chaîne phonétique produite par le phonétiseur du système de synthèse. La segmentation est ensuite effectuée sur la base de cette chaîne phonétique vérifiée. Là encore, le processus est en partie automatisé. La base étant annotée phonétiquement, il est alors possible d'apprendre des modèles acoustiques de type HMM, puis de les utiliser en vue d'obtenir une segmentation approchée. Ensuite, l'intervention d'opérateur humain se révèle indispensable pour affiner la position des frontières de phones. Signalons également que dans le cadre de la SPC, une annotation des unités est également nécessaire. Celle-ci consiste à déterminer les informations linguistiques (type de syllabe, position de la syllabe dans le mot, etc...) voire prosodiques (marqueurs mélodiques). Cette opération est menée automatiquement. Enfin, selon le mode de génération sonore du synthétiseur, d'autres traitements sont nécessaires. Par exemple, l'utilisation de l'algorithme TD-PSOLA suppose que les bases acoustiques soient marquées temporellement de façon pitch-synchrone. Pour cela, des algorithmes tels que [LC98] peuvent être utilisés. Au final, malgré les efforts considérables menés pour automatiser les traitements d'un enregistrement sonore, la création de voix reste un processus lourd et onéreux. Les tâches les plus coûteuses sont évidemment celles qui requièrent des vérifications intensives, à savoir la phonétisation et surtout la segmentation. Actuellement, environ deux mois sont nécessaires pour pouvoir disposer d'une

nouvelle voix de SPC opérationnelle. Ce processus de création de voix est certainement perfectible, notamment selon deux points principaux. D'une part, des progrès sont sans doute envisageables pour limiter les opérations de vérification, par exemple en utilisant des mesures de confiance pour détecter automatiquement les zones nécessitant une correction. D'autre part, des techniques de réduction de bases pourraient être incluses dans le processus de création de voix. Ce dernier aspect qui sera traité au chapitre 5 a une double utilité. La première est de limiter la taille des dictionnaires acoustiques dédiés à la SPC. Le deuxième avantage est qu'en opérant cette réduction de bases avant de procéder à la vérification de la segmentation, une diminution drastique de l'intervention humaine peut être obtenue.

2.3 Sélection des unités

Comme nous l'avons signalé précédemment, un système SPC repose sur l'utilisation de segments de signaux extraits de la parole naturelle et d'une algorithmique fine qui assurera la sélection des unités de la base les mieux adaptées au contexte. La méthode utilisée pour la sélection joue un rôle primordial pour ce type de synthèse. Nous présentons dans cette section le principe de la sélection des unités acoustiques dans un système SPC, ainsi que les paramètres utilisés pour une telle sélection.

2.3.1 Principe

La sélection consiste à déterminer la séquence d'unités acoustiques ayant les contextes de synthèse les mieux adaptés et minimisant les discontinuités aux instants de concaténation. Pour ce faire, les algorithmes de sélection visent à minimiser deux types de métriques : un "coût cible" qui mesure l'adéquation des unités avec les consignes (symboliques et/ou numériques) générées par les modules de traitements linguistiques du système et un "coût de concaténation" qui rend compte de la compatibilité acoustique et prosodique de deux unités consécutives. Plus précisément, la sélection consiste à minimiser une fonction de coût définie par :

$$C(u_1, ..., u_N) = \sum_{n=1}^{N} C_{cible}(u_n) + \alpha \sum_{n=2}^{N} C_{concat}(u_{n-1}, u_n), \tag{2.1}$$

où $C_{cible}(u_n)$ est le coût cible de l'unité u_n et $C_{concat}(u_{n-1}, u_n)$ le coût de concaténation entre les unités u_{n-1} et u_n, N désignant le nombre d'unités de synthèse. Dans cette équation, α est un facteur de pondération entre coût cible et coût de concaténation.

Un état de l'art relatif aux différents coûts cible et coûts de concaténation, ainsi que les différentes combinaisons sous forme de fonctions de coûts et leurs algorithmes de pondération seront détaillés dans la suite de cette section.

2.3.2 Coûts cible

Le coût cible est déterminé à partir de la comparaison entre les paramètres des unités candidates et ceux caractérisant la cible. Pour déterminer ces coûts cible, plusieurs auteurs utilisent des paramètres symboliques. Ainsi, Breen et Jackson dans [BJ98] calculent une distance entre une unité et la cible en utilisant des paramètres phonologiques dont les valeurs sont déterminées de façon manuelle par des séances d'essais/erreurs. De la même façon [PA01] utilisent des paramètres linguistiques (type de syllabe, position de la syllabe dans le mot, ...) pour la détermination des coûts cible. Ces méthodes présentent l'avantage d'être faciles à mettre en œuvre. En contrepartie, elles manquent de robustesse à cause de l'utilisation de tests informels qui dépendent de l'expert, peu représentatif d'un auditeur naïf.

Pour remédier aux problèmes de robustesse des séances d'essais/erreurs, certains auteurs déterminent leurs coûts cible (représentés par des différences de contexte phonémique) par le biais d'un test d'écoute formel [YG02]. Lee et Peng dans [LLO01] et [PZC02] utilisent, quant à eux, un test MOS (Mean Opinion Score) pour optimiser les valeurs des paramètres relatifs à des critères de ressemblance.

La majorité des systèmes actuels utilisent principalement des paramètres symboliques pour la sélection, d'une part, parce qu'il est difficile de définir des cibles numériques fiables et d'autre part, en raison de leur faible coût de calcul. Cependant, l'information symbolique n'englobe qu'une partie de la variabilité acoustique du signal de parole. Ainsi, pour éviter cet obstacle certains auteurs utilisent des paramètres acoustiques pour la détermination des coûts cible. Par exemple, Huang et al dans [HAH+97] comparent les fréquences fondamentales cibles et les fréquences fondamentales intrinsèques. De même, Donovan et al dans [DE98] comparent les durées cibles aux durées intrinsèques pour éliminer les unités ayant une durée trop éloignée de cette

cible. Iwahashi et al dans [IKS92] proposent, quant à eux, de minimiser la différence entre les caractéristiques spectrales cibles et les même caractéristiques intrinsèques de l'unité candidate.

D'autres auteurs utilisent des coûts mixtes ou des sous-coûts pour définir des coûts cible. Ainsi, Black et Campbell dans [BC95] utilisent à la fois des paramètres symboliques (tels que les traits articulatoires) et des paramètres acoustiques (durée, énergie, fréquence fondamentale). De même, Le Meur dans [Meu96] utilise un ensemble de sous-coûts (phonétique, morphologique et acoustique). L'algorithmie utilisée pour la détermination des poids attribués aux coûts sera présentée dans la section 2.3.4.

2.3.3 Coût de concaténation

Le coût de concaténation peut être déterminé à partir de la différence entre les paramètres acoustiques des deux unités consécutives, ou simplement par des informations symboliques. Cette dernière consiste à pénaliser certaines concaténations du type Consonne-Voyelle et Voisé-Voisé [Sag88]. Différemment mais dans le même objectif, Yi et Glass, dans [YG98], définissent, pour chaque classe phonémique à laquelle appartiennent les phones qui sont autour du point de concaténation, une valeur qui représente le degré de concaténation. Ces valeurs sont déterminées par le biais d'un test d'écoute formel.

La méthode la plus utilisée pour la détermination du coût de concaténation est basée sur une comparaison du signal de la fin de l'unité précédente avec celui du début de l'unité suivante. Ainsi, pour les points de concaténation situés dans des zones voisées, Hirokawa et Hakoda dans [Hir89] cherchent à minimiser la différence spectrale. Cependant, une distance spectrale faible au voisinage du point de concaténation n'implique pas forcément un signal résultant de bonne qualité [CI96]. Iwahashi et al. dans [IKS92] évaluent la distance acoustique entre le signal du contexte d'une unité et le signal de l'autre unité. Une autre façon de procéder est de prendre en compte plus précisément la dynamique du signal autour du point de concaténation. Certains auteurs utilisent les dérivées des coefficients acoustiques [BC95], [TKTS02]. Dans la même optique, Nomura et al dans [NMS90] comparent en trois endroits différents (unité précédente, point de concaténation et unité suivante) les paramètres LPC[1] et leurs vecteurs de moyennes

[1] Linear Predictive Coefficients

correspondants calculés sur toute la base de parole.

Pour rendre compte de la compatibilité de deux unités acoustiques il convient au préalable de définir une distance le plus en accord possible avec la perception de discontinuité par un auditeur humain. Une distance possible est la distance de Kullback-Leibler [KL51] calculée sur le spectre d'énergie et utilisée comme coût de concaténation par Klabbers et Veldhuis [KV98] dans leur système de synthèse. Hunt et Black dans [HB96] ainsi que Conkie et Isard dans [CI96] utilisent une distance euclidienne calculée entre les coefficients MFCC[2] de part et d'autre du point de concaténation. Donovan dans [Don01] propose une distance Mahalanobis entre les paramètres spectraux en employant un arbre de décision. Ces techniques tirent profit d'une bonne partie de l'information contenue dans le signal de parole, et représentent d'une façon numérique les discontinuités perceptibles. Cependant, après différents tests effectués sur les distances acoustiques les plus utilisées en synthèse, Stylianou et Syrdal dans [SS01] ont conclu que le pouvoir de prédiction de ces distances reste très limité. Ainsi la distance Kullback-Leibler parvient à détecter seulement 37% des discontinuités audibles. Ce score est faible et pousse à la recherche de nouvelles distances (ou combinaison de plusieurs distances) et nouveaux paramètres qui caractériseront plus précisément les discontinuités aux points de concaténation. Plus récemment, Pantazis et al dans [PSK05] ont pu augmenter le taux de prédiction à plus de 56% par la combinaison de deux nouveaux ensembles de paramètres. Le premier est calculé à partir d'une décomposition du signal de parole ; le second ensemble de paramètres est obtenu en utilisant une technique de séparation de l'amplitude et la fréquence dans un signal de parole. Les résultats de cette distance sont meilleurs par rapport aux autres distances, mais restent néanmoins faibles pour une prédiction précise.

2.3.4 Définition d'une fonction de coût

Une fonction de coût relative à une séquence d'unités est obtenue par combinaison des différentes fonctions de coût locales. Généralement une simple sommation de ces coûts locaux est effectuée, conformément à l'équation (2.1). Cependant, il est légitime de se demander si une telle stratégie est réellement adaptée. En effet, un des principaux défauts de la SPC est que des artefacts locaux peuvent apparaître. Ceux-ci sont dus

[2]Mel-Frequency Cepstral Coefficients

pour une large part au fait que dans le cas des coûts additifs de fortes discontinuités peuvent être tolérées si elles conduisent par ailleurs à une minimisation de la fonction de coût. Dans ce contexte, il semblerait donc préférable de davantage pénaliser ces fortes discontinuités. Pour ce faire, Toda et al. ont étudié dans [TKTS02] des fonctions de coût de la forme :

$$C_p(u_1, ..., u_N) = [(C_{cible}(u_n))^p + \sum_{n=2}^{N}(C_{concat}(u_{n-1}, u_n) + \alpha C_{cible}(u_n))^p]^{\frac{1}{p}}, \qquad (2.2)$$

pour p entier strictement positif. Le cas $p = 1$ correspond à l'équation (2.1), alors que le cas $p \to \infty$ correspond à la minimisation du coût local maximal sur l'ensemble de la séquence. D'après Toda et al. [TKTS02], il semble que le choix de $p = 2$ conduise au meilleur résultat, suivi d'assez près du cas $p = 1$ et de la mesure du coût local maximal.

Des coûts multiplicatifs ont également été proposés dans [HC98]. Mais dans ce cas, il est possible de se ramener à une simple addition par passage au logarithme. Ainsi, malgré les déficiences mentionnées ci-dessus, les fonctions de coûts basées sur la sommation de coûts locaux semblent être instaurées comme un standard.

Dans une fonction de coûts additive, il est rapidement apparu évident que tous les coûts locaux ne devaient pas être mis sur un même pied d'égalité, certains influençant plus que d'autres la qualité du résultat obtenu. Des recherches ont dès lors été réalisées afin de trouver la pondération idéale à appliquer lors du processus de sélection. Ainsi, Breen et Jackson dans [BJ98] utilisent une méthode experte pour la détermination des poids attribués à chaque coût. D'une autre façon, Black et Compbell dans [BC95] utilisent une régression linéaire basée sur une distance objective calculée en comparant le signal synthétique au signal naturel de plusieurs phrases. Utilisant la même distance objective, Le Meur dans [Meu96] optimise les poids des différents coûts avec un algorithme de recuit simulé pour converger vers une fonction de coût optimale.

2.4 Présélection des unités acoustiques

L'objectif de la sélection des unités est de choisir la séquence d'unités la mieux adaptée, c'est-à-dire qui conduira au signal de synthèse le plus proche possible de la

cible prédite et sans discontinuité acoustique audible. Pour des raisons de complexité algorithmique, énumérer et traiter d'emblée l'ensemble des combinaisons d'unités correspondant à la phonétisation d'un texte donné est difficilement envisageable. Il convient donc d'opérer un filtrage des données avant de décider du choix de la séquence optimale. Pour cette raison, le module de sélection opère généralement en deux étapes, illustrées à la figure 2.1 dans le cas d'utilisation de diphones.

- la première est qualifiée de " pré-sélection ", et consiste à sélectionner des ensembles d'unités pour chaque séquence cible. Ces unités, dites candidates, ont toutes un signal adapté pour la vocalisation du texte et sont donc toutes potentiellement sélectionnables ;
- la seconde étape est appelée " sélection finale " : elle retrouve parmi toutes les unités candidates la séquence d'unités optimale en minimisant une fonction de coûts préalablement définie. Ces deux étapes dépendent l'une de l'autre, mais peuvent toutefois être améliorées de manière indépendante, c'est pourquoi nous les présentons séparément dans ce qui suit. La partie sélection finale a été traitée en section (2.3.4).

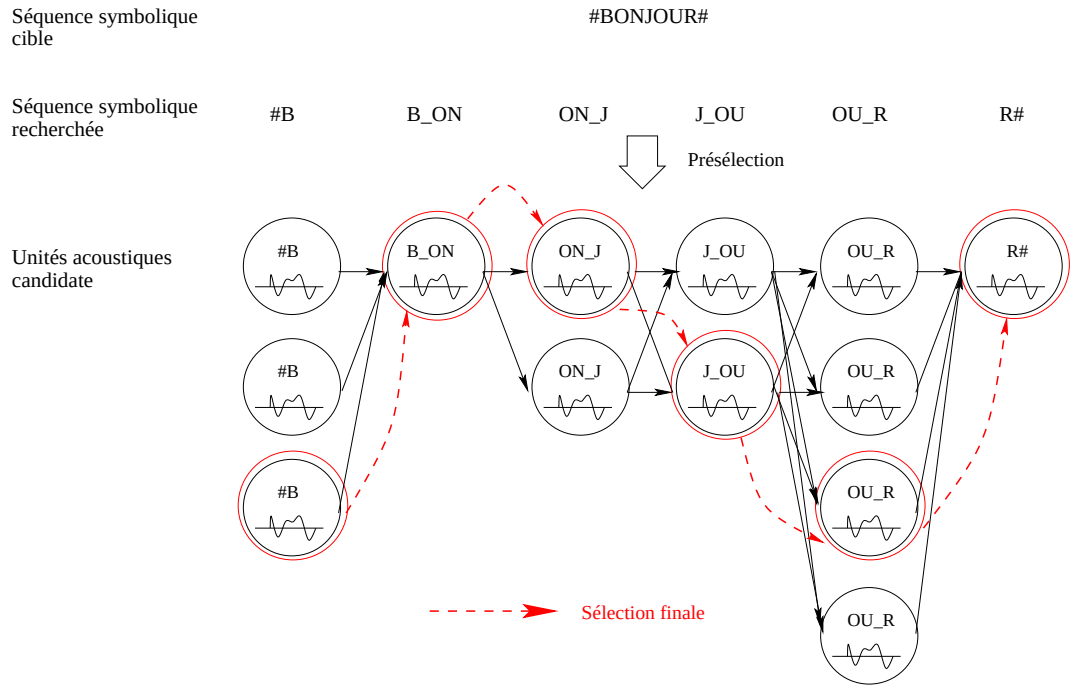


Figure 2.1 – Présélection et sélection finale des unités

2.4.1 Présélection par des méthodes à base d'expertise

L'objectif de la présélection est d'effectuer un filtrage des unités pour ne retenir que celles qui correspondent à la cible voire qui pourront être concaténées sans engendrer de discontinuité gênante. Cette étape de présélection est très délicate à mettre en œuvre, du fait de la multiplicité des critères entrant en jeu (phonémiques, phonologiques, prosodiques, etc...). Ainsi, le choix des paramètres symboliques pertinents est problématique, mais surtout l'importance relative de ces mêmes paramètres est très difficile à établir. Deux types de méthodes de présélection existent : d'une part les méthodes à base de règles définies de manière experte et d'autre part les méthodes automatiques qui seront présentées dans la section 2.4.2.

Les méthodes à base de connaissance d'expert consistent à définir un ensemble de règles destinées à localiser les unités sélectionnables pour la synthèse. Bien souvent des approches hiérarchiques sont utilisées. Ainsi Breen et al. [BJ98] commencent par rechercher les phonèmes ayant les meilleurs contextes phonémiques en parcourant un arbre d'indexation avant de réduire l'ensemble de recherche par le biais d'une fonction de coût. Citons également le système de synthèse de NTT [NH88] dans lequel la présélection se fait successivement sur des critères contextuels, prosodiques puis acoustiques (critère de continuité spectrale). Le système de synthèse du laboratoire ATR [SKIM92] est aussi basé sur une présélection hiérarchique. Hiérarchiser les critères présente l'avantage de réduire l'espace de recherche, mais l'inconvénient de privilégier de manière systématique certains critères par rapport à d'autres. Une autre manière de procéder consiste à combiner linéairement les différents critères. Mais se pose alors le problème du choix des coefficients de pondération associés aux différents critères. Ces coefficients sont définis à partir de connaissances d'expert, mais nécessitent en général un important travail d'ajustement manuel. Un autre défaut de ces méthodes est qu'elles peuvent être dépendantes de la base acoustique. Cela signifie que pour utiliser une nouvelle base de données acoustiques, il faut vérifier que les règles et coefficients définis conviennent et, le cas échéant, les réajuster manuellement.

2.4.2 Méthodes automatiques

Vu les difficultés de mise en œuvre des algorithmes de présélection et le manque de flexibilité de la sélection finale des méthodes basées sur des connaissances d'expert,

l'automatisation du processus de sélection (présélection et sélection finale) s'est avéré inévitable. Au début, la communauté scientifique s'est focalisée sur l'automatisation de l'étape de présélection en s'inspirant de la méthode COC (Context Oriented Clustering) proposée par Nakajima dans [NH88]. Puis, elle a essayé de la généraliser en l'utilisant pour la classification et la sélection de modèles de Markov cachés (HMM), dans un contexte de synthèse de parole par corpus. Une présentation plus détaillée des modèles HMM est fournie en annexe.

2.4.2.1 Context Oriented Clustering (COC) [NH88]

La méthode COC consiste à générer une partition des unités au moyen d'un arbre de décision. La création d'un tel arbre est un processus récursif où l'on cherche, pour chaque nœud, à effectuer une séparation des données à partir d'un paramètre symbolique. Le choix du paramètre symbolique utilisé et de la valeur de séparation associée est réalisé de façon à optimiser un critère de variance acoustique des ensembles d'unités des nœuds fils. La figure 2.2 montre un exemple d'arbre obtenu par la méthode COC.

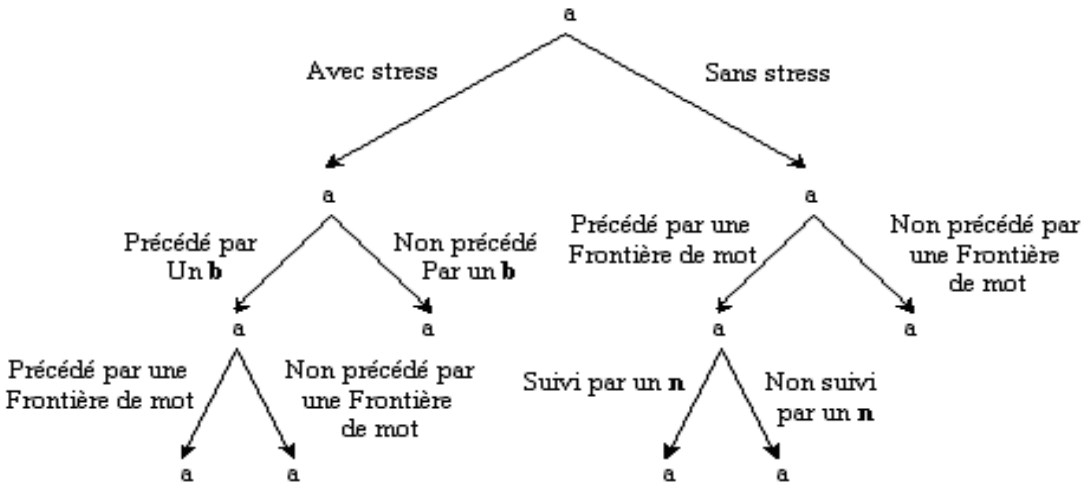


Figure 2.2 – Exemple d'arbre de décision généré par la méthode COC pour l'unité "a"

Originellement, les unités présentes dans les feuilles servent ensuite à générer des unités " moyennes ", appelées centroïdes des unités de la feuille, qui seront utilisées

pour la synthèse vocale [NH88]. Il convient néanmoins de préciser que cette méthode ne fut pas conçue à l'origine à des fins de présélection. En effet, l'arbre ainsi généré n'est pas utilisé pour la suite des traitements, mais à la synthèse, une fonction définie par les auteurs est appliquée sur les centroïdes pour réaliser la sélection proprement dite.

2.4.2.2 Variantes de la méthode COC

Plusieurs auteurs ont ensuite repris et légèrement modifié cet algorithme dans un contexte de synthèse de parole. Ainsi, Wangt et al dans [WCI93] ont utilisé l'arbre de décision généré pour prédire les unités (phones) à sélectionner pour la synthèse de l'anglais. De plus, ils y ont ajouté une procédure de validation croisée dans le but d'optimiser la taille de l'arbre. Cette approche a ensuite été reprise par Black dans [BT97] et a donné naissance à un module de présélection opérationnel intégré dans le système de synthèse Festival. La méthode COC a également été appliquée pour la classification de modèles HMM. En reconnaissance de la parole, Bahl et al dans [BSG+91] ont utilisé des arbres de décision afin d'obtenir une classification en sénones[3] en contexte. Donovan dans [Don96] et Huang dans [HAH+97] ont ensuite repris cette méthode, l'ont adaptée au contexte de synthèse de parole afin de réaliser un module de présélection intégré dans le synthétiseur d'IBM dans le cas de Donovan et de Microsoft dans le cas de Huang.

Citons également les travaux de Tokuda et al., qui, dans [TZB02], proposent une méthode de synthèse basée sur le même type de classification acoustique. Néanmoins, le but de cette classification n'est pas d'opérer une présélection, mais plutôt de définir une cible acoustique utilisée à des fins de génération. Pour cela ils utilisent un algorithme de synthèse basé sur un modèle source-filtre [FKI92]. La qualité obtenue par un tel synthétiseur est cependant médiocre, et ceci essentiellement parce qu'il n'existe pas de technique de synthèse capable de générer un signal de parole paraissant naturel uniquement à partir d'informations acoustiques (coefficient MFCC par exemple).

[3] Un sénone est un segment acoustique correspondant à un état d'un modèle HMM

2.5 Conclusion

Ce chapitre a permis de présenter les principales caractéristiques de la SPC. Nous avons pu voir que la qualité de la parole produite par un système de SPC est étroitement liée d'une part au corpus de synthèse et d'autre part à l'algorithmie de sélection. Malgré le saut notable de qualité qu'a permis d'atteindre cette technologie, la SPC se heurte à un problème de robustesse, c'est-à-dire qu'elle n'est pas capable de garantir une parole dont la qualité soit à peu près constante sur l'ensemble d'un énoncé. Ceci a d'ailleurs été un frein considérable au développement de SPC. À titre d'illustration, en 1998, une compagne d'évaluation menée dans le cadre d'un Workshop dédiée à la synthèse vocale [WTTS98], a fait ressortir que la synthèse à base de polysons sélectionnés statistiquement (technique développée conjointement par les BellLabs et Lucent Technologies) est de qualité meilleure que le système de SPC CHATR proposé par ATR. Le manque de robustesse de la SPC est en partie dû à une couverture acoustique du corpus limitée, mais aussi et à notre sens surtout, au manque de contrôle acoustique des systèmes de SPC actuels. En effet, un des défauts de la plupart des systèmes de SPC est que les critères de sélection font essentiellement intervenir des coûts cible de nature symbolique. Or, ceux-ci sont au mieux capables de donner une caractérisation moyenne de certains corrélats acoustiques (fréquence fondamentale, MFCC, etc...). En revanche, la diversité acoustique des unités ayant des descripteurs symboliques similaires n'est absolument pas prise en compte dans le coût cible. Certes, des coûts de concaténation ont été définis pour tenter de maîtriser cette diversité acoustique, mais, comme nous l'avons souligné en section 2.3.3, leur pouvoir de prédiction de discontinuités audibles reste relativement limité. Il semble par conséquent qu'un meilleur contrôle de l'acoustique produite par un système de SPC passe par une meilleure maîtrise des fonctions de coût. Pour cela, des coûts plus pertinents doivent être proposés et une algorithmie plus fine doit être mise en œuvre pour pénaliser toute discontinuité potentielle. Dans le cadre de cette thèse, nous nous intéressons essentiellement à la prise en compte de cibles acoustiques dans le processus de sélection. L'hypothèse sous-jacente est que l'utilisation de cibles acoustiques judicieuses devrait fournir une mesure acoustique plus fine de l'adéquation d'une unité candidate au contexte de synthèse que ne le ferait une mesure définie sur un espace discret. Le respect d'une telle cible devrait alors permettre de mieux maîtriser la variabilité acoustique des unités candidates. C'est dans cette optique que nous présentons une nouvelle méthode de sélection au chapitre 3.

Chapitre 3

Sélection des unités par le biais d'une cible acoustique

3.1 Introduction

Ce chapitre est consacré à l'introduction de critères acoustiques pour la sélection des unités. Plus précisément, des informations spectrales sont définies et considérées comme des consignes acoustiques à respecter. La prise en compte de ce type d'informations dans le processus de sélection lui-même est relativement marginale. Au chapitre précédent, nous avons fait état de plusieurs techniques de classification acoustique des unités. Ce type de classification permet de séparer les unités acoustiquement en fonction de leurs caractéristiques symboliques, de sorte qu'à la synthèse des classes d'unités relativement homogènes puissent être aisément localisées sur la base de seuls critères symboliques, par exemple via le parcours d'arbres de décision. Les fonctions de coûts cible se limitent donc à une juxtaposition de critères symboliques et l'information acoustique des unités n'est pas explicitement exploitée dans la définition du coût cible. Seule la compatibilité acoustique de deux unités est prise en compte, via un critère de distorsion spectrale évalué lors de la concaténation de deux unités. Dans ce chapitre, nous cherchons donc à prendre en compte explicitement l'information acoustique des unités en définissant un coût cible acoustique. Avant d'aller plus loin, il convient cependant de noter que si des coûts cible acoustiques n'ont pas été intégrés dans le processus de sélection lui-même, la nature acoustique des unités a tout de même été prise en compte

en vue d'effectuer un filtrage de ces unités. Ainsi, la méthode COC, telle qu'originellement proposée dans [NH88], effectue à la fois une classification et une sélection d'unités candidates sur la base d'informations acoustiques. Plus précisément, au sein de chaque classe acoustique, seule l'unité acoustique la plus proche du centroïde spectral de la classe considérée est conservée. Donovan reprend ce principe pour déterminer un ensemble d'unités utilisables dans le cadre de la synthèse par sénones [Don96]. De manière similaire dans [HPAA99] Huang et al. proposent une méthode visant à ne retenir pour la synthèse que les diphones les plus pertinents au sens d'une mesure acoustique. Pour ce faire, ils construisent des modèles HMM de diphones en contexte à 4 états à partir des modèles de sénones. Un tel modèle de diphone est donc caractérisé par 4 modèles de sénones, ce qui permet de définir pour chaque instance de diphone une mesure de vraisemblance. La réduction des unités est ensuite faite en considérant les instances de diphones pour lesquelles les modèles ayant les mêmes sénones initiaux et finaux et en ne retenant que l'instance ayant la vraisemblance la plus élevée. Les méthodes mentionnées ci-dessus peuvent donc être considérées comme des méthodes de pré-sélection ou encore de réduction de dictionnaire acoustique. Dans le cadre de cette thèse, nous souhaitons conserver une mesure spectrale cible au sein même du processus de sélection des unités. A ce titre, les méthodes présentées au paragraphe précédent peuvent être utilisées, car elles permettent de définir un coût cible à chaque unité. Ce dernier peut ensuite être combiné à un coût de concaténation en vue de la détermination d'une séquence d'unités optimale au sens du critère de sélection ainsi défini. Cependant, un coût cible basé sur un critère de vraisemblance tel qu'employé par exemple dans [Don96] s'avère en pratique assez limité, dans la mesure où les modèles HMM sont essentiellement réputés pour leurs capacités à caractériser les zones stables des phonèmes. En revanche ces modèles, lorsqu'ils sont utilisés à des fins de décodage acoustico-phonétique, ne sont pas capables de modéliser finement des trajectoires acoustiques. Le fait qu'ils ne prennent pas en compte les dépendances temporelles entre les observation, les rend peu apte à modéliser finement les trajectoires acoustiques. De ce fait, le score de vraisemblance obtenu sur les parties transitoires des phonèmes reste peu fiable et par conséquent une approche basée sur la maximisation de la vraisemblance vis-à-vis de modèles HMM semble peu adaptée à la qualification des zones transitoires. Une autre façon de procéder est d'utiliser les modèles HMM à des fins de génération de trajectoires acoustiques. En effet, disposant d'une séquence de modèles de sénones auxquels sont associées des densités gaussiennes, il devient possible d'inférer des trajectoires acoustiques, par exemple conformément à la

méthode présentée dans [MTKI96]. La séquence de paramètres acoustiques ainsi obtenue peut ensuite être utilisée pour générer de la parole synthétique. Cette technique de synthèse de la parole, baptisée synthèse par HMM et décrite dans [TZB02], ne permet pas la restitution d'une parole synthétique de très haute qualité, mais les trajectoires spectrales ainsi générées sont considérées comme satisfaisantes et permettent d'obtenir une assez bonne intelligibilité. C'est sur la base de ce constat que nous nous proposons d'exploiter la cible ainsi produite à des fins de sélection des unités. Le principe de la méthode est présenté en 3.2. La phase d'apprentissage des modèles acoustiques est détaillée en section 3.3 alors que le système de synthèse basé sur une sélection acoustique est présenté en 3.4. La section 3.5 décrit les expérimentations que nous avons menées afin d'évaluer la méthode proposée.

3.2 Principe de la méthode

Le principe de la méthode de synthèse proposée est explicité à la figure 3.1. Il consiste à effectuer une sélection sur la base d'une cible purement acoustique, qui est générée à partir des paramètres HMM. La méthode proposée est constituée de deux phases principales. Tout d'abord, un apprentissage est mené sur une base de données et conduit à une classification acoustique des unités (sénones) en fonction de critères symboliques (contexte phonétique, informations linguistique et prosodique, etc...). Durant la synthèse, les modules haut-niveau du système effectuent l'analyse linguistique et prosodique du texte à synthétiser. Les paramètres symboliques cible obtenus permettent alors, à partir de la classification préalablement réalisée, de déterminer la séquence de modèles acoustiques adaptée au contexte de synthèse. Les paramètres de ces modèles sont ensuite utilisés pour définir une cible acoustique. Enfin, un module de sélection vise à déterminer la séquence d'unités la plus proche de cette cible acoustique.

3.3 Apprentissage des modèles acoustiques

Cette section présente la première partie de la méthode (réalisée hors ligne), qui consiste à préparer les modèles acoustiques. Pour cela, une modélisation acoustique par HMM est employée. Comme nous l'avons souligné dans le chapitre précédent, l'utilisation des modèles HMMs nécessite deux étapes : une étape d'apprentissage au cours

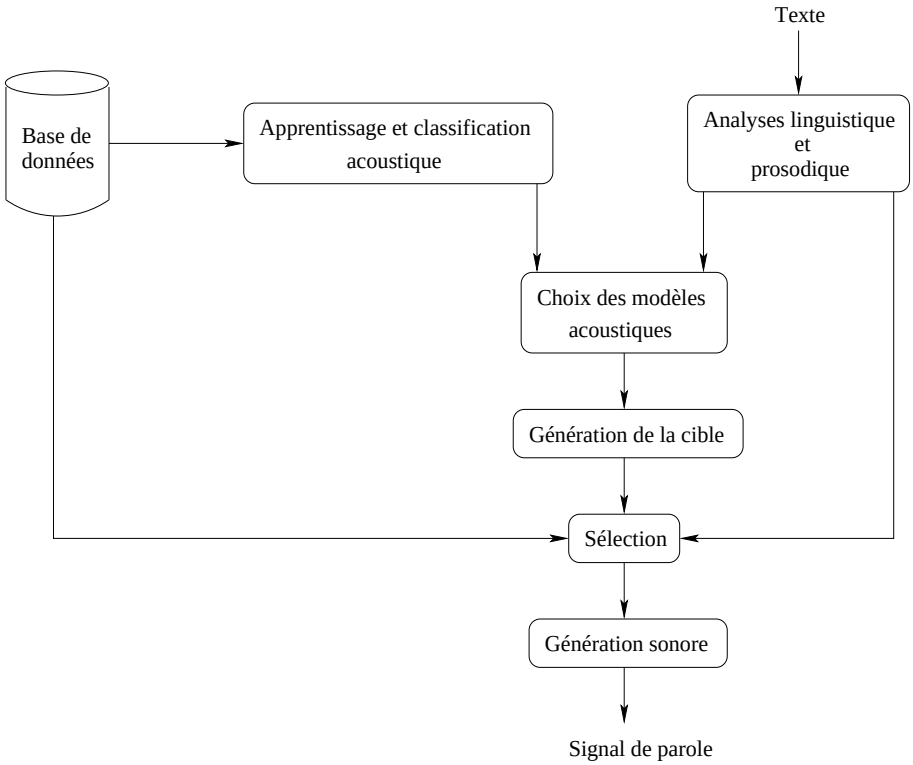


Figure 3.1 – Architecture générale de la méthode proposée

de laquelle le processus stochastique est estimé à partir d'observations extensives via l'algorithme de Baum-Welch et une étape de mise en oeuvre où le modèle peut être utilisé en temps réel à des fins de décodage, c'est-à-dire pour obtenir la séquence d'états de modèles la plus probable, cette dernière étant obtenue par l'algorithme de Viterbi.

La Figure 3.2 représente les différentes étapes de la classification acoustique. Dans un premier temps des modèles HMMs sont appris pour les différents phonèmes. Ensuite tous les triphones contenus dans la base de données acoustiques sont identifiés et un modèle est estimé pour chacun d'eux. Enfin une classification est opérée via la construction d'un arbre de décision afin de regrouper au sein d'une même classe des modèles acoustiquement proches. Dans la suite de cette section, ces différentes étapes sont détaillées.

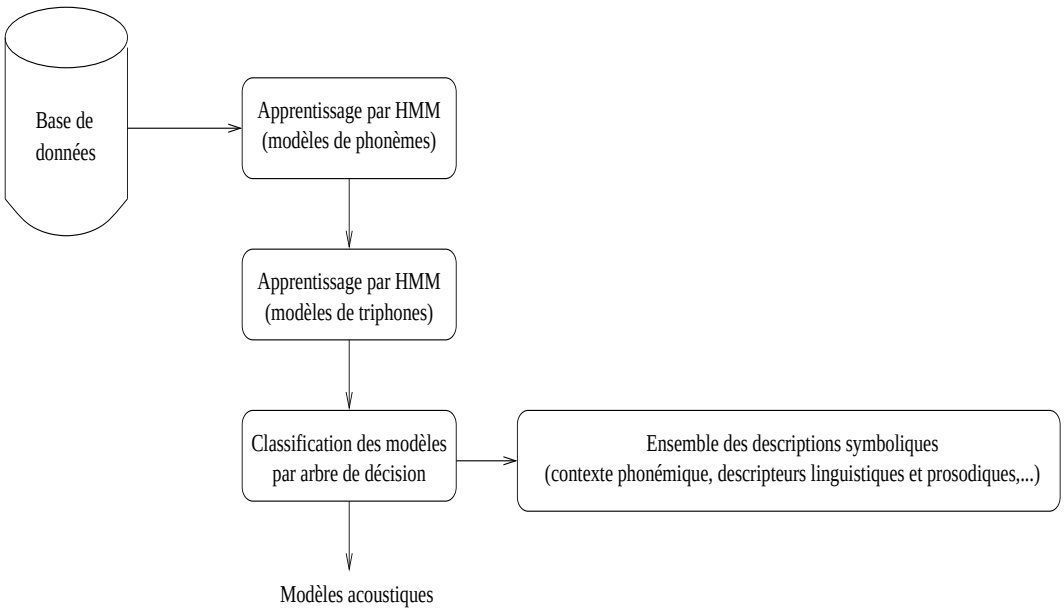


Figure 3.2 – Les différentes étapes de la classification acoustique.

3.3.1 Apprentissage des modèles de phonème

Afin d'effectuer l'apprentissage des modèles, nous supposons que nous disposons d'une base acoustique contenant le signal de parole proprement dit et la transcription phonétique qui lui est associée. Nous nous plaçons donc dans le cadre d'un apprentissage supervisé. L'analyse acoustique est menée d'une manière asynchrone avec un pas fixe de 5 ms. Le signal est analysé avec une fenêtre de 10 ms et pour chaque fenêtre, 12 MFCC (Mel Frequency Cepstral Coefficients) [DM80] et l'énergie [TFBH94] sont extraits. Le but final de l'extraction des paramètres est de modéliser la parole, un phénomène très variable. Par exemple, même si elle a de l'importance, la simple valeur de l'énergie n'est pas suffisante pour donner toute l'information portée par ce paramètre. Il est donc souvent nécessaire de recourir à des informations sur l'évolution dans le temps de ces paramètres. Cette information de la dynamique temporelle du signal est exploitable en utilisant, outre les paramètres statiques, les paramètres différentiels de ces paramètres statiques [Fur86]. Pour cela, les dérivées première et seconde sont calculées pour représenter la variation ainsi que l'accélération de chacun des paramètres. Même si la robustesse de la représentation obtenue est accrue, cela implique aussi de multiplier par 3 l'espace de représentation. Le calcul de la dérivée première se fait par

la formule suivante :

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}, \tag{3.1}$$

où d_t est la dérivée à l'instant t calculée à partir des coefficients MFCCs statiques $c_{t-\Theta}$ à $c_{t+\Theta}$. Le paramètre Θ permet le lissage temporel de ces coefficients dérivées [Fur86]. La valeur de Θ, dans la littérature, varie de 1 à 6. Dans ce travail, nous fixons la valeur de ce paramètre à 2. Pour la dérivée seconde la formule est appliquée en changeant les coefficients MFCCs par les valeurs de la dérivée première.

A chaque phonème est associé un HMM gauche-droite à trois états et pour chaque état du modèle, une loi gaussienne de moyenne μ et de covariance diagonale Σ modélise la distribution des observations. L'apprentissage de chacun de ces modèles est réalisé de manière classique par l'algorithme de Baum-Welch. Les différentes étapes de l'apprentissage des modèles de phonèmes sont représentées sur la Figure 3.3.

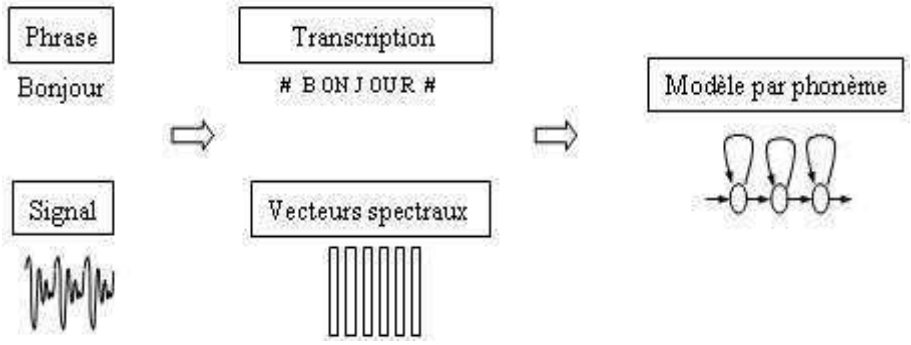


Figure 3.3 – Les différentes étapes pour l'apprentissage des modèles HMM par phonème.

3.3.2 Apprentissage des modèles de triphone

Les phonèmes représentent en phonologie le découpage des mots en sous-unités linguistiques. Un phone désigne quant à lui une réalisation acoustique d'un phonème. Or, les réalisations acoustiques des phonèmes sont différentes suivant le contexte d'élocution. Par exemple, selon le contexte phonétique du phonème considéré, des phénomènes de coarticulation sont observés de manière plus ou moins importante. De plus, en fonction du contexte prosodique, des différences de réalisation acoustique peuvent également apparaître. Il est alors judicieux de caractériser ces différences acoustiques et donc d'affiner

les HMM en fonction du contexte. La majorité des systèmes tient compte uniquement des contextes phonémiques gauche et droit, ce qui aboutit à une modélisation dite par triphone. Lors de l'apprentissage de tels modèles, pour chaque triphone présent dans la base, les paramètres des lois gaussiennes relatives à chaque état du triphone sont réestimés à partir des représentants de ce triphone. Lorsque le nombre de représentants d'un triphone dans le corpus acoustique est insuffisant, les paramètres du modèle de ce triphone risquent d'être mal estimés. Il est cependant possible, en regroupant les phonèmes des contextes gauche et droit en classes, d'obtenir des modèles plus génériques dépendant du contexte. A titre d'exemple, les contextes peuvent être regroupés en classes telles que les plosives, les fricatives, les liquides, ou encore les classes voisées ou non voisées.

3.3.3 Classification par arbres de décision

Après apprentissage des modèles de triphones, une classification est effectuée, afin de regrouper au sein d'une même classe des modèles acoustiquement proches. Une telle classification peut être obtenue par la construction d'arbres de décision. Un arbre de décision est construit pour chaque état de chaque phonème. Cette construction est réalisée par divisions répétées des données en sous-ensembles, ces divisions étant opérées sur les paramètres symboliques. A chaque nœud de l'arbre, une question du type "le phonème à droite est-il voisé ?" est posée [BSG+91]. En plus des questions contextuelles, d'autres questions sur les marqueurs mélodiques et les positionnements syllabiques ont été utilisées dans ce travail pour leur très forte discrimination acoustique. La liste des différentes questions utilisées est fournie en annexe.

Pour chaque question, une séparation des observations contenues dans le nœud père est effectuée et la variation de vraisemblance entre le nœud père et les deux nœuds fils est estimée [Ode95]. Ce calcul de vraisemblance est réalisé à partir des paramètres des modèles de triphones déterminés précédemment. La question conduisant à l'augmentation maximale de la vraisemblance est retenue et la séparation est effectivement acceptée si cette augmentation de vraisemblance dépasse un seuil fixé et si le nombre de représentants présents dans chacun des nœuds fils est suffisant. À chaque fois que la séparation est acceptée, les paramètres des nouveaux nœuds sont réestimés. Il est à noter que seule la moyenne μ et la matrice de covariance Σ seront réestimées sans la matrice des transitions. D'après Odell [Ode95], les variations de la matrice de transi-

tion ne sont pas significatives et peuvent être considérées comme constantes durant la classification.

Cette opération de séparation est répétée sur chaque branche, jusqu'à ce qu'un critère d'arrêt stoppe le partitionnement et donne lieu à la génération d'une feuille de l'arbre. Cette classification est effectuée une seule fois lors de la conception du synthétiseur. La Figure 3.4 montre un exemple d'arbre de décision. Les nœuds terminaux finaux appelés aussi feuilles forment les états groupés pour chaque arbre. A chacune des feuilles de l'arbre est associée une loi gaussienne de moyenne μ_c et matrice de covariance Σ_c caractérisant les représentants de cette feuille.

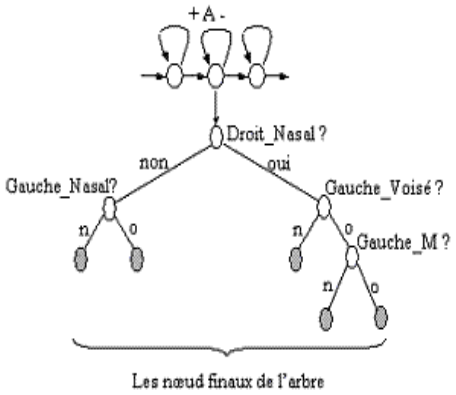


Figure 3.4 – Exemple d'arbre de décision pour le deuxième état du phonème "A".

La séparation des sénones issus d'un nœud de l'arbre est effectuée si les deux conditions suivantes sont remplies :

- le nombre de sénones sur chacun des nœuds fils est supérieur à un nombre minimal n_{min} ;
- l'augmentation de la vraisemblance provoquée par la séparation dépasse un certain seuil S.

La détermination de ces deux critères est délicate car il convient de trouver un compromis entre le pouvoir discriminant du classifiant et sa capacité de généralisation. Si les seuils n_{min} et S sont trop bas, il risque de se produire un sur-apprentissage. Le corollaire est que les contextes non vus dans le corpus seront mal modélisés. Au contraire, des seuils trop élevés rendraient cette classification peu judicieuse car peu

discriminante.

Ces deux critères d'arrêt sont fixés par une procédure de validation croisée [CM98]. Les résultats de la validation croisée seront présentés et commentés en section 3.5.

3.4 Mise en œuvre de la méthode proposée

La mise en œuvre de la méthode proposée passe par la présentation du système de synthèse sur lequel la méthode est appliquée. Le système de synthèse proposé est présenté à la figure 3.5 Il est constitué de cinq blocs de traitements principaux. Tout d'abord, une analyse linguistique et prosodique du texte à synthétiser est réalisée et permet de déterminer, parmi l'ensemble des classes acoustiques, la séquence de modèles acoustiques adéquate. Puis, à partir de cette séquence de modèles acoustiques et d'une information de durée, une cible acoustique est générée par le biais d'un algorithme de lissage. L'étape suivante est le découpage de la cible résultante en segments dont la taille dépend de la nature des unités utilisées à la synthèse (phones, diphones, demi-phones, etc). L'opération de sélection proprement dite est constituée d'une présélection sur des paramètres linguistiques et prosodiques suivie d'une sélection finale. Cette dernière est réalisée en calculant une distance acoustique entre chaque unité candidate et les segments cibles précédemment définis. Enfin, le signal de parole synthétique est obtenu par concaténation des instances sélectionnées. Les différentes étapes concernant la mise en œuvre de la méthode proposée sont détaillées dans les paragraphes suivants.

3.4.1 Analyse du texte et recherche des modèles acoustiques

Après une première étape d'analyse linguistique qui a pour but d'extraire du texte à synthétiser les informations symboliques (linguistiques et prosodiques), une recherche de la séquence de modèles acoustiques est menée dans l'ensemble des classes acoustiques. Lors de cette recherche, les arbres de classification correspondant aux éléments cibles sont parcourus afin de trouver la séquence de feuilles dont les paramètres symboliques sont les plus proches de ceux de la cible.

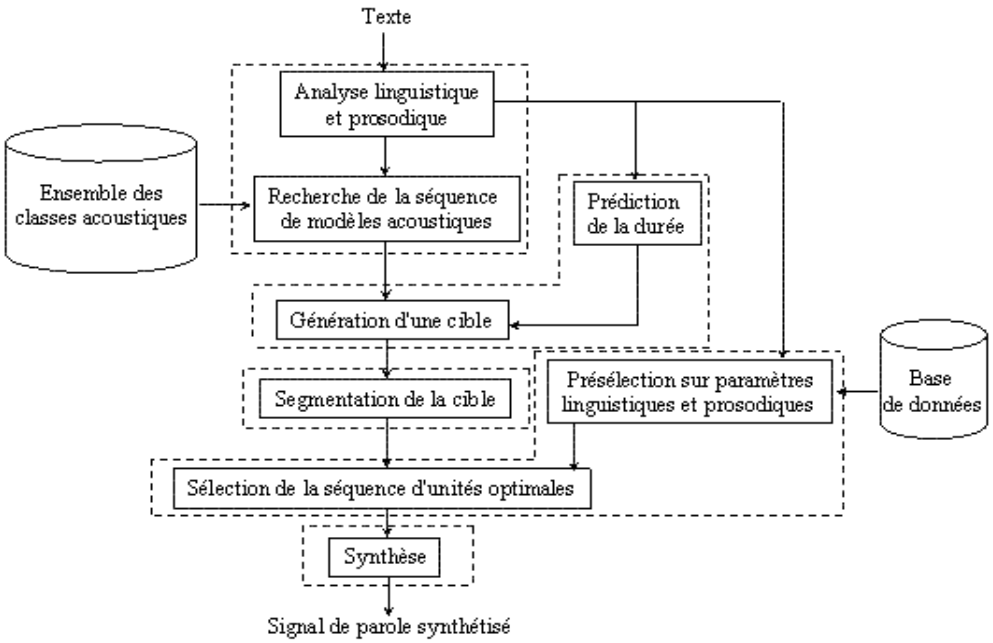


Figure 3.5 – Architecture du système de synthèse proposé.

3.4.2 Génération de la cible acoustique

3.4.2.1 Détermination de la durée

L'algorithme de génération de la cible acoustique prend en entrée une séquence de paramètres de modèles HMM (moyenne et covariance des densités gaussiennes pour chaque état et matrice de transition) et une valeur de durée pour chaque état des modèles. Cette durée doit bien entendu être proche de la durée moyenne des sénones en contexte considéré. Ici, nous ne faisons pas de prédiction de la durée en fonction du contexte, mais nous nous limitons à la durée moyenne de chaque état calculée sur toute la base acoustique. Bien entendu d'autre modèles de prédiction de durée plus performants peuvent être utilisés pour assigner une durée aux différents états des modèles acoustiques sélectionnés. Les modèles de prédiction de durée existant dans la littérature visent à attribuer à chaque phonème une valeur de durée. Parmi ceux-ci on peut citer la méthode présentée dans [Pie80] et celle basée sur le concept d'élasticité des syllabes présentée par [CI91]. A partir de chaque consigne de durée phonémique d, il convient de déterminer des durées pour chaque état de ce phonème. Pour cela, il est nécessaire de calculer, pour chaque modèle λ , la durée de chaque état i (notée α_i^λ), donnée par

la relation suivante :

$$\alpha_i^\lambda = \frac{1}{1 - a_{ii}}, \tag{3.2}$$

où a_{ii} est la probabilité *a priori* de rester dans l'état i. La durée de l'état i du modèle λ considéré est alors

$$d_i^\lambda = \frac{\alpha_i^\lambda}{\sum_{i=1}^{I} \alpha_i^\lambda} d. \tag{3.3}$$

Connaissant cette valeur d_i^λ, il est alors aisé de déterminer le nombre de trames de l'état i pour le modèle λ considéré.

3.4.2.2 Détermination de la cible

Ayant déterminé une séquence de modèles acoustiques et une durée relative à chaque état du modèle, il convient maintenant de générer une cible adéquate. Soient N le nombre total de trames à synthétiser, $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_N]$ la séquence des modèles acoustiques cible et $Q = [q_1, q_2, \ldots, q_N]$, la séquence d'états correspondante. Le but est alors de déterminer la séquence d'observations $O = [o_1^T, o_2^T, \ldots, o_N^T]^T$ maximisant $P[O|Q, \Lambda]$. Dans notre cas, le vecteur d'observation o_t de la trame t est constitué d'une partie statique $c_t = [c_t(1), c_t(2), \ldots, c_t(p)]^T$ (les p coefficients MFCC) et d'une partie dynamique $\triangle c_t, \triangle^2 c_t$ (la dérivée première et la dérivée seconde des coefficients MFCC), d'où $o_t = [c_t^T, \triangle c_t^T, \triangle^2 c_t^T]^T$ avec :

$$\triangle c_t = \sum_{i=-L^{(1)}}^{L^{(1)}} w^{(1)}(i) c_{t+i}, \tag{3.4}$$

$$\triangle^2 c_t = \sum_{i=-L^{(2)}}^{L^{(2)}} w^{(2)}(i) c_{t+i}. \tag{3.5}$$

La séquence d'observation s'écrit aussi sous forme matricielle de la façon suivante :

$$O = WC, \tag{3.6}$$

avec

$$C = [c_1, c_2, \ldots, c_N]^T, \tag{3.7}$$

$$W = [w_1, w_2, \ldots, w_N]^T, \tag{3.8}$$

$$w_t = [w_t^0, w_t^1, w_t^2] \tag{3.9}$$

et

$$\begin{aligned} w_t^{(n)} = [&0_{P\times P}, \ldots, 0_{P\times P}, w^{(n)}(-L^{(n)})I_{P\times P}, \\ &\ldots, w(n)(0)I_{P\times P}, \ldots, w(n)(L^{(n)})I_{P\times P}, \\ &0_{P\times P}, \ldots, 0_{P\times P}]^T, \qquad n = 0, 1, 2. \end{aligned} \tag{3.10}$$

Maximiser $P[O|Q, \Lambda]$ par rapport à O revient à résoudre

$$\frac{\partial \log P(O|Q, \Lambda)}{\partial C} = 0, \tag{3.11}$$

avec

$$\log P(O|Q, \Lambda) = -\frac{1}{2}O^T U^{-1} O + O^T U^{-1} M + K, \tag{3.12}$$

$$U^{-1} = diag[U_{q_1}^{-1}, U_{q_2}^{-1}, \ldots, U_{q_N}^{-1},], \tag{3.13}$$

et

$$M = [\mu_{q_1}^T, \mu_{q_2}^T, \ldots, \mu_{q_N}^T]^T. \tag{3.14}$$

où μ_{q_t} est le vecteur moyenne et U_{q_t} la matrice de covariance de l'état q_t, K étant une constante indépendante du vecteur d'observation O. L'équation (3.11) devient :

$$RC = r, \tag{3.15}$$

avec

$$R = W^T U^{-1} W \tag{3.16}$$

et

$$r = W^T U^{-1} M^T. \tag{3.17}$$

Comme R est une matrice de $(NP \times NP)$ éléments, la résolution directe de l'équation (3.15) nécessite N^3P^3 opérations. Pour réduire la complexité de l'algorithme, une procédure itérative de lissage telle que celle proposée par [MTKI96] peut être employée. Cette procédure consiste à calculer les coefficients statiques à l'instant t à partir seulement des coefficients statiques et dynamiques des instants $t-1$ et $t+1$. Nous avons essayé d'appliquer cette procédure en considérant le signal des sénones adjacents. La figure 3.6 présente une comparaison des trois méthodes : la résolution directe, la méthode de [MTKI96] dite RLS[1] et la méthode [MTKI96] appliquée par état de modèle HMM

[1]Recursive Least Squares

dite RLS par état. Le temps est présenté sur l'axe des abscisses et la différence entre les coefficients statiques naturels et ceux estimés par les trois méthodes sur l'axe des ordonnées .

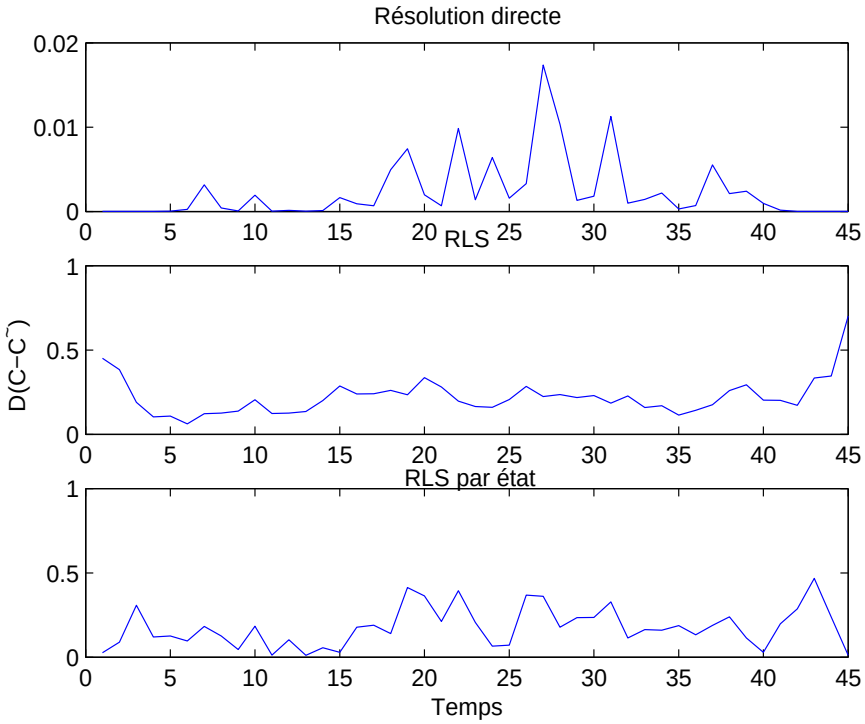


Figure 3.6 – Comparaison des trois méthodes d'estimation des coefficients statiques : résolution directe, RLS et RLS par état.

Comme présenté dans la figure 3.6, les coefficients estimés par la résolution directe sont les plus proches des coefficients naturels. Cependant, pour une phrase d'une longueur moyenne (comme celle de l'exemple) la résolution directe prend plus d'une heure pour estimer tous les coefficients de la phrase. Les méthodes RLS et RLS par état sont en temps réel. En moyenne, les deux méthodes RLS et RLS par état sont équivalentes, sauf aux frontières de la phrase où la méthode RLS par état est meilleure que la méthode RLS car elle est plus proche à la résolution directe.

3.4.3 Segmentation de la cible

La Figure 3.7 illustre la génération de la cible acoustique pour le mot "BONJOUR". Pour pouvoir utiliser cette cible acoustique à des fins de sélection, il est nécessaire d'opérer un découpage de cette cible en fonction des types d'unités de synthèse souhaités (diphones, phones, demi-phones, etc...).

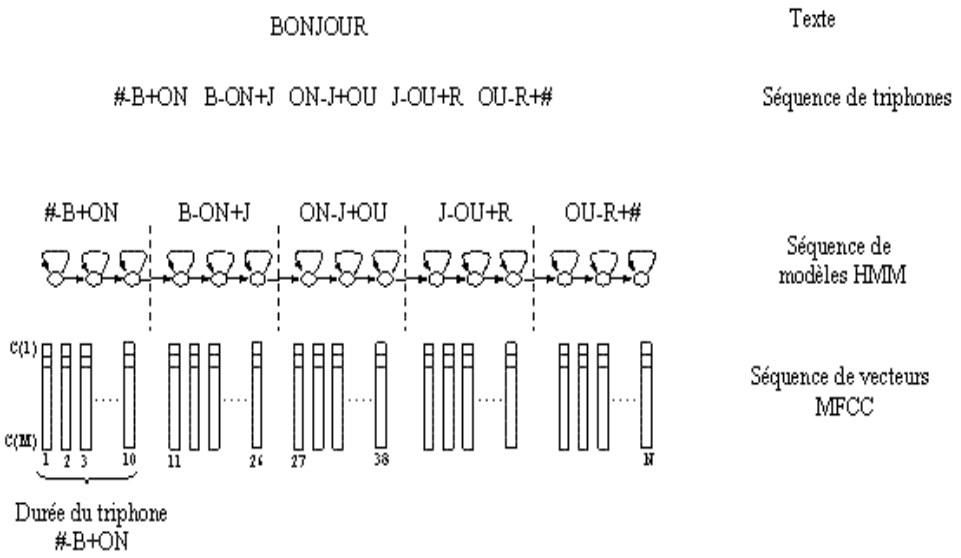


Figure 3.7 – Exemple de génération d'une cible acoustique.

Notons d'ailleurs que l'un des points forts de la méthode proposée est qu'elle est applicable à tout type de synthèse (par diphones, demi-phones, etc...), la seule opération à effectuer étant de segmenter la cible pour l'adapter aux unités de la base. Par exemple, pour passer d'une segmentation par phonèmes à une autre par diphones, nous prendrons les trames de la deuxième moitié du premier phonème du diphone et la première moitié du phonème qui le suit pour former une unité diphone, comme explicité à la Figure 3.8.

3.4.4 Sélection de la séquence d'unités optimales

Nous décrivons ici comment la cible obtenue précédemment peut être utilisée à des fins de sélection des unités. Dans ce travail, nous ne faisons état que de la synthèse

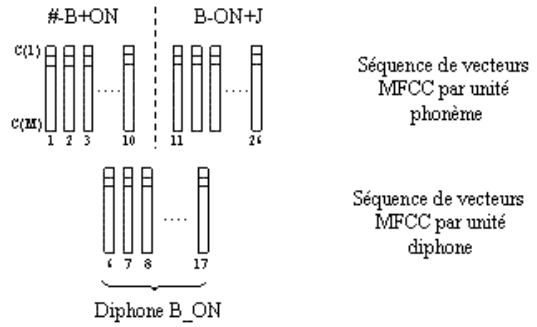


Figure 3.8 – Passage d'une segmentation en phonèmes à une segmentation en diphones.

par diphones, étant entendu que la même méthodologie peut être employée pour tout autre type d'unités de synthèse. Chaque unité de type diphone est représentée par un nombre variable d'instances. A chaque instance est associée une séquence de vecteurs MFCC, issue de l'étape d'analyse acoustique.

Pour chaque diphone à synthétiser, les instances de ce diphone sont comparées au segment de la cible correspondant à ce diphone par le biais d'un algorithme de DTW (Dynamic Time Warping). Cet algorithme de DTW effectue un alignement des deux segments et permet en outre de calculer une distance globale entre ces derniers, égale à la somme des distances locales sur le chemin d'alignement pondérée par un terme sanctionnant une différence de durée relative entre les signaux comparés (3.18). Si par exemple on compare une unité cible a et une des unités candidate b représentées comme deux séquences de vecteurs acoustiques, ayant respectivement N_a et N_b éléments, les différences locales entre leurs vecteurs sont calculées par une distance d, euclidienne en l'occurrence. Ainsi, la distance locale entre le vecteur i de l'unité cible a le vecteur j de l'unité candidate b vaut d_{ij}. La distance globale D entre les deux unités est calculée selon la formule :

$$D = \frac{max(N_a, N_b)}{min(N_a, N_b)} \times \sum_{chemin} d_{ij} \tag{3.18}$$

Ainsi, les informations permettant de retrouver dans la base acoustique les différentes unités sélectionnées sont transmises au module de synthèse. La génération sonore consiste simplement à récupérer dans la base de données les instances sélectionnées

et à les concaténer. Une opération de lissage peut être utilisée lors de la synthèse, afin de minimiser les discontinuités (de spectre, de pitch ou d'énergie) aux instants de concaténation.

3.5 Expérimentations

Cette section présente les tests menés afin de fixer les critères d'arrêts des arbres de décision et d'évaluer la procédure de sélection proposée. Premièrement, la préparation des données et le détail de toutes les étapes d'apprentissage seront présentés. Puis, un premier test sera réalisé et discuté pour justifier le choix des deux critères d'arrêts. Le deuxième test s'appuie sur les résultats issus du premier pour effectuer une évaluation de la méthode proposée. L'évaluation est conduite en comparant, par un test MOS (Mean opinion score), la procédure de sélection proposée à celle utilisée dans le système de synthèse de FTR&D.

3.5.1 Apprentissage des modèles et classification

La base de données utilisée pour fixer les critères d'arrêt et évaluer la méthode de sélection proposée au cours de cette étude a été développée au sein de l'équipe de synthèse de FTR&D. Cette base contient 7 heures et demie de parole générée par un locuteur. Les paramètres acoustiques et linguistiques qui caractérisent le signal et les textes lus sont aussi inclus dans la base.

Le passage du signal de parole étiqueté vers un modèle HMM par triphone nécessite plusieurs étapes. Premièrement, les paramètres spectraux (MFCC, énergie) sont extraits du signal ainsi que leurs dérivées premières et secondes. Ensuite, un modèle HMM gauche droite de trois états est associé à chaque allophone. Chaque modèle doit être appris : les moyennes, les variances et les probabilités de transition entre états sont réestimées jusqu'à ce qu'un seuil de convergence ou qu'un nombre maximum d'itérations soit atteint. Ceci est fait par l'algorithme de Baum-Welch. Enfin, un autre étiquetage de la base, basé sur le contexte gauche et droit du phonème ainsi que son marqueur mélodique et sa position syllabique, est réalisé. Les modèles acoustiques d'allophones appris précédemment sont réestimés de nouveau sur la base de ce nouvel étiquetage pour devenir en final des modèles de triphones. Ces étapes sont plus détaillées dans

[Don96]. Pour la réalisation de ces différentes étapes un environnement HTK[2] a été utilisé. Plusieurs outils de ce dernier ont été reprogrammés pour permettre, en plus des informations contextuelles, d'utiliser des informations prosodiques durant les étapes d'estimation des modèles de triphone et de classification de ces modèles.

En effectuant quelques tests informels nous avons remarqué que l'énergie et sa dérivée première et seconde ont peu d'influence sur le choix de l'unité acoustique par notre méthode. Pour confirmer cette remarque, nous avons effectué un apprentissage et une classification par des vecteurs qui contiennent l'énergie et sa dérivée première et seconde et un apprentissage et une classification par des vecteurs sans l'énergie et sa dérivée première et seconde. Les arbres de décision résultant des deux classifications étaient presque identiques. Il semble donc que l'influence de l'énergie soit assez faible lorsqu'il s'agit de faire ressortir des différences liées aux contextes phonémiques. En revanche, dans une tâche de reconnaissance de la parole, l'information d'énergie reste pertinente, car elle peut permettre de discriminer entre plusieurs modèles de phonèmes. Dans la suite de cette étude, nous avons décidé de ne pas considérer l'énergie dans les vecteurs acoustiques.

Pour fixer les deux critères d'arrêt des arbres de décision (nombre minimum de candidats par feuille et maximum de vraisemblance), un test de validation croisée est mené. Pour cela, la base de données a été découpée en K sous-bases. Ce test consiste, pour plusieurs couples de critères d'arrêts à tester, à réaliser la partie apprentissage sur un ensemble de $(K-1)$ bases et de considérer la base restante comme base de test. Pour chaque phrase de la base de test considérée, une cible acoustique est générée puis comparée à la phrase naturelle via un algorithme de DTW. Cette procédure permet de calculer des distances DTW moyennes pour chaque couple de critères d'arrêt testé. Ces distances sont présentées à la figure 3.9.

On observe sur la figure 3.9 que les distances DTW associées aux différents critères d'arrêts sont relativement proches. Cependant, il s'avère qu'une faible variation de la distance DTW a une forte influence sur le choix des unités sélectionnées et donc sur la qualité de la parole synthétisée.

Par ailleurs, un compromis qualité de classification / complexité d'exploration d'arbre est à rechercher dans la détermination des valeurs des critères d'arrêts. Pour le premier

[2]Hidden Markov Model Toolkit

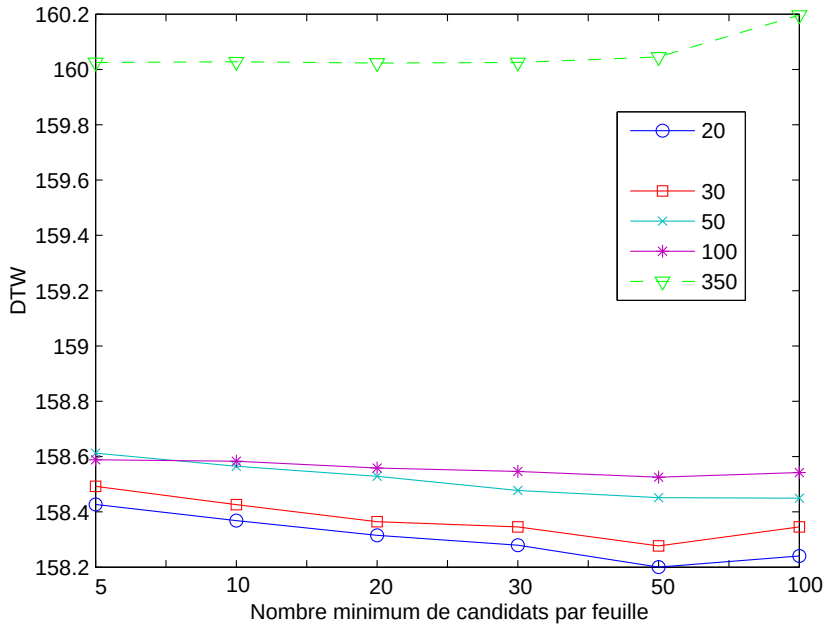


Figure 3.9 – Résultats de la validation croisée des critères d'arrêts en fonction du nombre de candidats par feuille et du seuil d'augmentation minimum de la vraisemblance

critère (nombre de candidats par feuille), on observe dans la figure 3.9 qu'un minimum de 50 candidats par feuille est la valeur optimale. Par contre, pour le deuxième critère (maximum de vraisemblance), durant le test on a remarqué que les faibles valeurs de vraisemblances augmentent la complexité de l'arbre pour un gain en qualité minime. Pour cela, le deuxième critère d'arrêt a été fixé par inspiration des arbres résultants. Nous avons choisi une configuration conduisant à un nombre de feuilles moyen par arbre d'environ 80.

3.5.2 Évaluation de la méthode de sélection proposée

Dans cette section nous évaluons la qualité globale de la synthèse obtenue avec deux méthodes de sélection, celle présentée dans cette étude et celle utilisée dans le système de synthèse de FTR&D [Blo03]. La méthode de sélection proposée, appelée dans cette évaluation "HMM", utilise les critères d'arrêts des arbres de décision issus des résultats du premier test. La deuxième méthode testée, dite "FTR&D", utilise, pour la présélection des n-diphones à synthétiser, une heuristique basée sur quelques sous-

coûts symboliques. La sélection finale de la méthode FTR&D utilise une optimisation d'une fonction de coûts additive (symboliques et acoustiques) par un algorithme de type Viterbi [Vit67]. Il est à noter que les deux méthodes testées n'utilisent, durant la concaténation, aucun algorithme de lissage de la fréquence fondamentale ou de l'énergie.

Pour ce test, 20 phrases phonémiquement équilibrées tirées des ensembles définis par [Com81] ont été utilisées. Les sujets qui ont participé à ce test, au nombre de 17, sont tous naïfs par rapport à la synthèse vocale, de langue maternelle française et sans problèmes d'audition reconnus.

Lors du test, les 20 phrases sont vocalisées avec chacune des deux configurations de sélection testées puis présentées aux sujets dans un ordre aléatoire, chaque phrase n'étant présentée qu'une seule fois. L'écoute des stimuli est réalisée avec un casque audio professionnel dans un bureau calme, les sujets pouvant ajuster le niveau d'écoute à leur convenance. Après avoir écouté une phrase, les sujets doivent noter sa qualité globale sur une échelle à cinq niveaux, répertoriés dans le tableau 3.1.

5	Excellente
4	Bonne
3	Moyenne
2	Médiocre
1	Mauvaise

Tableau 3.1 – Les notations possible dans un test MOS

Afin que les sujets puissent étaler leur notation sur toute l'échelle de valeurs, une session d'apprentissage est dispensée avant le test. Cet apprentissage consiste à leur présenter 4 stimuli, couvrant l'étendue de la gamme de qualité des phrases synthétisées du test, sans avoir à les noter. Chacune des deux versions de sélection testées est présentée par 2 phrases d'apprentissage. Afin d'évaluer la cohérence de chaque sujet, une phrase par configuration de sélection est présentée deux fois aux sujets. Ces phrases en double servent d'indicateur de cohérence : seule leur note moyenne par sujet est utilisée dans le calcul du MOS. Ainsi, ces paires de phrases de contrôle ont la même contribution au MOS que chacune des autres phrases.

3.5.2.1 Résultats

Tout d'abord, deux sujets parmi les 17 participants à ce test ont attribué des notes différant de deux unités aux deux phrases de contrôles. Ceci nous conduit à l'éviction des résultats de ces deux sujets. Pour les 15 sujets restants, il s'avère, après un examen préliminaire des notes, que chaque sujet a utilisé toute la gamme de notation allant de un à cinq, ce qui tend à valider la session d'apprentissage menée au début du test. En plus, les sujets semblent tous avoir un comportement similaire. En effet, les répartitions des notations sont à peu près homogènes.

Les résultats des MOS des deux méthodes sont présentés sur le tableau 3.2. Le MOS attribué à la configuration de sélection de FTR&D est supérieur à celui attribué à la méthode proposée. Cependant, il est remarquable de constater que l'utilisation d'un seul coût acoustique conduise à ce degré de qualité.

Méthode	**MOS**
FTR&D	3,53
HMM	3,09

Tableau 3.2 – Résultats des tests MOS

Les phrases pour lesquelles la différence des notes MOS dépasse une unité sont au nombre de 5 sur les 19 du test (sans la phrase du contrôle) comme le montre la figure 3.10. Le défaut principal perçu, lors d'écoutes informelles par des experts de ces 5 phrases vocalisées par la méthode proposée (HMM), est la présence de discontinuité au niveau de la fréquence fondamentale. Cette discontinuité s'explique par le fait qu'à aucun moment, que ce soit lors de la classification ou lors de la sélection elle-même, la fréquence fondamentale n'est prise en compte. Il n'est donc pas étonnant que ce paramètre soit mal contrôlé lors de la synthèse.

En revanche pour deux phrases la synthèse par la méthode proposée s'est montrée significativement meilleure que celle obtenue par la méthode FTR&D. Une écoute de ces phrases a signalé des problèmes de discontinuités spectrales audibles lorsque le système de synthèse de FTR&D été utilisé, ces discontinuités étant absentes avec la méthode proposée. Ceci étant, le bilan global est tout de même en faveur de la méthode FTR&D.

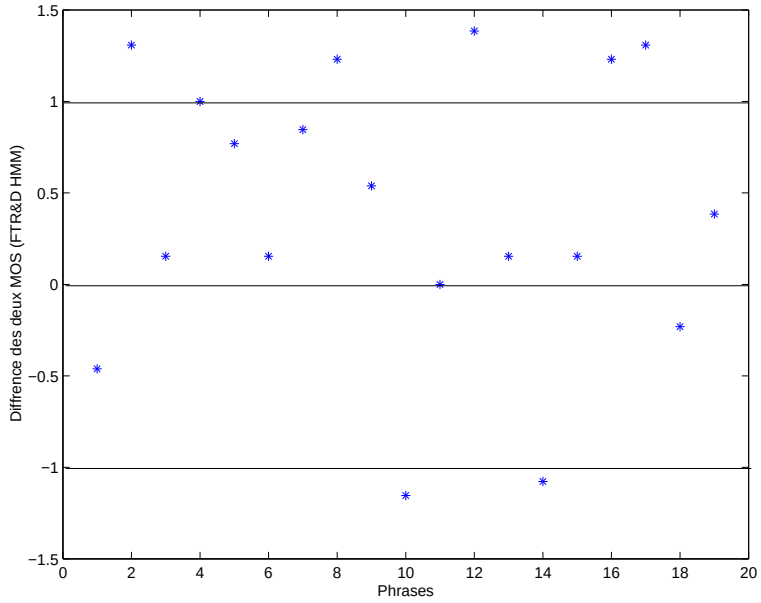


Figure 3.10 – Différence des MOS attribués par phrase

3.6 Conclusion

Dans ce chapitre nous avons proposé une nouvelle méthode de sélection basée sur la génération d'une cible acoustique. Les différentes étapes nécessaires à la mise en œuvre ont été ainsi présentées. La première d'entre elles concerne l'apprentissage des modèles HMM de chaque phonème puis de chaque triphone existant dans le corpus. Ensuite, une classification par arbre de décision est opérée et aboutit à un ensemble de modèles de sénones caractérisés selon leurs contextes linguistique et prosodique. Lors de la synthèse, les informations symboliques déterminées lors des traitements linguistiques du texte en entrée permettent de définir de manière non ambiguë la séquence de modèles nécessaire à la génération d'une cible acoustique. La sélection consiste alors à rechercher la séquence d'unités minimisant la distance à cette cible.

L'évaluation de la méthode a été réalisée, comparativement à la méthode de sélection actuelle du système de synthèse de FTR&D, par un test d'écoute formel. Globalement, une préférence est donnée à la méthode de FTR&D. Une analyse plus fine des résultats a fait ressentir qu'une des déficiences de la méthode proposée était le manque de contrôle de la fréquence fondamentale. Cependant, un des atouts de la méthode proposée est qu'elle impose des contraintes acoustiques fortes, ce qui a pour effet de limiter les

discontinuités spectrales. En résumé, ce chapitre a permis de montrer le potentiel de la méthode proposée tout en soulignant certaines de ses déficiences. Des marges de progrès ont été identifiées, notamment en ce qui concerne un meilleur contrôle de la fréquence fondamentale. Pour les satisfaire, il apparaît nécessaire d'enrichir notre méthode de sélection actuelle. Ces modifications font l'objet du chapitre suivant.

Chapitre 4

Perfectionnement de la méthode proposée

4.1 Introduction

Lors de l'évaluation menée au chapitre précédent, nous avons souligné que la méthode proposée ne permettait pas d'avoir un contrôle suffisant de la fréquence fondamentale. Ceci se manifeste à la synthèse par la présence d'artefacts dus à des discontinuités de fréquence fondamentale. L'effet de ces discontinuités peut certes être réduit en effectuant un lissage de la fréquence fondamentale via des algorithmes de modification prosodiques classiques tels que TD-PSOLA ou HMM, mais ces algorithmes sont eux-mêmes susceptibles de dégrader la qualité du signal de parole. Nous proposons donc dans ce chapitre une modification de la procédure de sélection visant à mieux contrôler la fréquence fondamentale. Dans un premier temps, nous prenons explicitement en compte la fréquence fondamentale dans le processus de classification en introduisant cette information dans les vecteurs acoustiques. En outre, nous proposons un processus de sélection en deux étapes : une étape de présélection dont le but est d'isoler un ensemble restreint d'unités candidates sur la base de la distance acoustique définie au troisième chapitre et une étape de sélection finale dans laquelle un coût de concaténation sanctionnant les différences de la fréquence fondamentale est introduit. Un autre objectif de ce chapitre est de limiter la complexité du module de sélection. En effet, telle qu'exposée au troisième chapitre, la méthode proposée nécessite des calculs de distance

entre la cible acoustique générée et toutes les unités candidates. Or, il semble clair que de nombreuses unités ont des caractéristiques acoustiques peu compatibles avec cette cible, d'où la nécessité d'un filtrage de ces unités indésirables. Un tel filtrage doit être peu complexe pour être efficace et doit donc reposer sur des informations symboliques. Dans ce chapitre, nous proposons de réutiliser le module de présélection de FTR&D.

4.2 Prise en compte de la fréquence fondamentale

Cette section présente tout d'abord l'algorithme d'estimation de la fréquence fondamentale utilisée dans cette thèse. Une mise en forme des vecteurs acoustiques et l'apprentissage des nouveaux modèles HMM sont ensuite présentés et justifiés. Des tests sont effectués et leurs résultats discutés à la fin de cette section.

4.2.1 Estimation de la fréquence fondamentale

L'estimation de la fréquence fondamentale est un problème difficile de l'analyse de la parole. Elle est également un des plus importants, du fait de la très grande sensibilité de l'oreille à la fréquence fondamentale [Hes83]. L'estimation est rendue difficile par plusieurs facteurs :

- le signal de parole, considéré stationnaire sur une trame de courte durée (10 à 30 ms), présente souvent des ruptures. Dans les régions de transition, les caractéristiques de la parole peuvent changer rapidement.
- la possibilité d'une présence simultanée d'une excitation voisée et non-voisée.
- l'éventail des fréquences fondamentales est assez large, ces dernières pouvant varier de 50Hz pour des voix masculines particulièrement graves à 400Hz pour des voix d'enfant.

Pour répondre à ces problèmes, plusieurs algorithmes ont été proposés [MQ90] [Her88] [GH87]. Dans cette thèse, l'algorithme d'estimation du pitch utilisé est celui proposé par Stylianou [Sty96]. Cet algorithme consiste à estimer, dans une première étape, la fréquence fondamentale $\hat{f}_0$ par une méthode temporelle basée sur la maximisation de la fonction d'autocorélation. Ensuite, ce pitch initial est utilisé pour la détection du voisement puis affiné par une méthode fréquentielle.

4.2.2 Prise en compte du pitch dans le vecteur acoustique

Dans la littérature, la modélisation de la fréquence fondamentale est généralement réalisée par des modèles statistiques de type HMM. L'application des modèles HMMs ordinaires (à l'aide de la loi d'émission mono ou multi-gaussienne) pour la modélisation de la fréquence fondamentale est confrontée à un sérieux problème dû à la non-définition de cette dernière (F_0) dans les parties non-voisées du signal. Autrement dit, la séquence d'observation du modèle de la fréquence fondamentale est composée de valeurs continues à une dimension et de symbole discret qui représente le voisement. Pour remédier à ce problème plusieurs méthodes ont été proposées pour contrôler la partie non-voisée du signal. Ainsi, [FF88] met à la place de chaque partie non-voisée un vecteur aléatoire généré à partir d'une fonction de densité de probabilité avec un très grand écart type, et modélisant ensuite ces vecteurs aléatoires explicitement dans le modèle HMM. Dans la même optique, [RO94] suppose que les valeurs de la fréquence fondamentale existent dans les parties non-voisées du signal sauf qu'elles ne sont pas observables.

À l'opposé, d'autres auteurs (comme [TMMK99]) cherchent plutôt à jouer sur le modèle HMM en lui même. Pour cela, ils proposent un nouveau type de modèle HMM pour la modélisation de la fréquence fondamentale, dans lequel les probabilités d'émission d'un état sont définies par les distributions de probabilité de type MSD-HMM[1]. L'objectif de cette modélisation est de pouvoir inclure, dans le même modèle, une modélisation HMM discrète et une autre continue avec mélange de gaussiennes. L'avantage d'une telle modélisation est que les vecteurs d'observation peuvent être de différentes dimensions, ce qui permet d'inclure le vecteur d'ordre zéro qui représente l'observation d'une partie non-voisée.

Dans le cadre de cette thèse, le vecteur des paramètres acoustiques utilisé pour l'apprentissage des modèles HMM, comme présenté dans la section 3.2.1, est composé des 12 coefficients MFCC, de l'énergie et de leurs dérivées première et seconde. Pour les raisons évoquées dans le chapitre précédent l'énergie a été supprimée du vecteur acoustique, son emplacement dans le vecteur a été utilisé pour y mettre la valeur de la fréquence fondamentale qui correspond à ce vecteur.

Dans le cas de phonèmes voisés, l'approche utilisée dans ce travail consiste à combiner chaque valeur de la fréquence fondamentale avec le vecteur MFCC correspondant,

[1] Multi-Space probability Distribution HMM

et de modéliser leur densité de probabilité par des modèles HMM. Avant d'insérer ces valeurs dans les vecteurs acoustiques correspondants, une normalisation de ces valeurs de fréquence fondamentale est effectuée selon l'équation :

$$F_{log} = \log(\frac{F_0}{\bar{F_0}}), \tag{4.1}$$

où F_0 est la fréquence fondamentale en Hz et $\bar{F_0}$ est la moyenne des valeurs de pitch sur tout le corpus.

Pour la dérivée première et seconde de la fréquence fondamentale nous avons utilisé l'équation (3.1) qui sera dans ce cas :

$$\triangle f_t = \frac{\sum_{\theta=1}^{\Theta} \theta(f_{t+\theta} - f_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}, \tag{4.2}$$

où $\triangle f_t$ est la dérivée à l'instant t calculée à partir des valeurs du pitch $f_{t-\Theta}$ à $f_{t+\Theta}$. Dans ce travail la valeur de Θ est fixée à 2. Pour la dérivée seconde, la formule est appliquée en changeant les valeurs de la fréquence fondamentale par les valeurs de la dérivée première. La figure 4.1 présente les paramètres acoustiques d'un phonème voisé.

Figure 4.1 – Disposition des paramètres acoustiques d'une trame voisée.

Dans le cas où le phonème est non voisé ou semi-voisé (voisé que dans certains contextes par exemple le [R]), les modèles HMM correspondants sont ceux présentés dans le chapitre précédent, c'est-à-dire que les vecteurs acoustiques utilisés pour l'apprentissage sont composés par les 12 MFCC et leurs dérivées premières et secondes.

Pour différencier la notation de ces nouveaux vecteurs par rapport à ceux présentés dans le chapitre précédent, nous noterons dans la suite de ce document les anciens vecteurs par :

– MFCC_D_A : (12 MFCC) + (12 $\triangle$MFCC) + (12 $\triangle^2$MFCC).

et les nouveaux par :

– MFCC_F_0_D_A : (12 MFCC + F_0) + (12 $\triangle$MFCC + $\triangle F_0$) + (12 $\triangle^2$MFCC + $\triangle^2 F_0$) pour les phonèmes voisés.

– MFCC_D_A1 : (12 MFCC) + (12 $\triangle$MFCC) + (12 $\triangle^2$MFCC) pour les phonèmes non voisés.

4.2.3 Apprentissage et classification des nouveaux modèles

Comme dans le chapitre précédent, un premier apprentissage est réalisé sur des modèles HMM gauche-droite associés à chaque phonème. Puis, ces modèles sont raffinés en fonction de leurs contextes phonémiques gauche et droite et donneront naissance à des modèles de triphones. Ces derniers sont classifiés par la construction d'arbres de décision. Les mêmes questions que celles utilisées au chapitre précédent sont posées pour la construction de ces arbres et les mêmes critères d'arrêt sont utilisés pour stopper le partitionnement de ces arbres. Le tableau 4.1 présente (pour les deux analyses MFCC_F_0_D_A / MFCC_D_A1 et MFCC_D_A) le nombre de nœuds terminaux par état pour tous les phonèmes.

État HMM / Type d'analyse	État 1	État 2	État 3
MFCC_D_A	357.7	594.94	376.08
MFCC_F_0_D_A et MFCC_D_A1	325.58	641.85	336.94

Tableau 4.1 – Moyenne de feuilles par état pour tous les phonèmes.

4.2.4 Sélection des unités

Durant la synthèse, une recherche des séquences des modèles acoustiques est menée en parcourant les arbres de classification en fonction des informations contextuelles du texte à synthétiser. Une cible acoustique est générée à partir de ces modèles puis segmentée en unités de synthèse correspondantes. Il est à noter que seuls les MFCC sont générés. La génération de la fréquence fondamentale est omise et cela pour de multiples raisons. La première cause est due à la présence des zones de transitions voisées vers non voisées et l'inverse. Les fréquences fondamentales générées dans ces zones seront forcément erronées. Les autres raisons seront présentées plus loin.

Ensuite, une présélection acoustique est réalisée sur la base de la cible ainsi segmentée. Cette présélection consiste à comparer, pour chaque diphone à synthétiser,

les instances de ce diphone au segment de la cible correspondant par le biais d'un algorithme de DTW et de garder les N meilleures instances les plus proches du segment cible. Il est à noter également que les distances locales calculées par l'algorithme DTW ne tiennent compte que des paramètres spectraux (MFCC), ceci est pour éviter l'influence des erreurs de segmentation et le fait que la pondération entre ces deux informations (MFCC et fréquence fondamentale) est difficile à ajuster.

Finalement, parmi les N meilleures instances de tous les diphones, la séquence d'unités qui assure une bonne continuité de pitch est sélectionnée. Pour cela, un algorithme de Viterbi est utilisé pour la sélection de cette séquence d'unités optimale en minimisant une fonction de coût. Dans le cas où la concaténation est effectuée sur un phonème voisé, le coût utilisé est composé de la différence de pitch entre les deux unités à sélectionner et une distance entre les paramètres spectraux de ces même unités. Par contre, si la concaténation est réalisée sur un phonème non voisé, le coût utilisé est simplement une distance entre les paramètres spectraux des deux unités à sélectionner. Cette distance, proposée par [Don01], est donnée par :

$$D_{MFCC} = \sqrt{\sum_{i=1}^{N} [\frac{e_i - \mu_i^P}{\sigma_i^P}]^2}, \tag{4.3}$$

où e_i est la différence entre les vecteurs MFCC d'ordre N et μ_i^P et σ_i^P sont respectivement la $i^{ème}$ composante du vecteur des moyennes et de la matrice de covariance du vecteur e pour le phonème P où la concaténation aura lieu. μ^P et σ^P sont calculés sur toute la base d'apprentissage.

Le choix de cette configuration est validé par les différents tests objectifs présentés dans la section 4.2.5.2.

4.2.5 Expériences et résultats

Tout d'abord, après plusieurs tests, nous avons remarqué que les instances sélectionnées dans la séquence optimale se trouvent à chaque fois parmi les 20 premières instances présélectionnées. Pour cela, le nombre des N meilleures instances à pré-sélectionner a été fixé à 20.

Dans cette section, nous présentons les résultats des expérimentations que nous

avons effectuées en vue de l'évaluation de la prise en compte du pitch dans le cadre de l'amélioration de la méthode proposée.

Dans ces expérimentations, nous avons évalué subjectivement ansi qu'objectivement trois méthodes de synthèse. Il s'agit pour deux d'entre elles de la méthode proposée au chapitre 3 et dans ce chapitre, alors que la troisième est la version actuellement commercialisée du système de synthèse de France Télécom (appelée dans cette expérimentation FTR&D).

Pour cette évaluation, la méthode proposée dans le chapitre 3 a été modifiée. Elle contient des modules de pré-sélection et sélection finale telle que présentés en 4.2.4 de façon à pouvoir être comparée à la méthode proposée dans ce chapitre. En effet, dans ce cadre, les deux configurations ne diffèrent que par la classification acoustique. Ainsi, la nouvelle configuration de la méthode du chapitre précédent est appelée dans cette expérimentation HMM_MFCC1 et celle présentée dans ce chapitre est appelée HMM_MFCC_F_0.

4.2.5.1 Evaluation subjective

Tout comme le test subjectif du chapitre précédent, ce test est lui aussi mené selon les recommandations de la norme P800 ACR. Cette fois, 19 sujets (tous naïfs par rapport à la synthèse vocale, de langue maternelle française et sans problèmes d'audition connus) ont participé à ce test. Les stimuli utilisés consistent cette fois en 23 phrases phonémiquement équilibrées issues du corpus de test défini par [Com81]. Ces phrases sont synthétisées avec chacune des trois méthodes de sélection testées, puis présentées aux sujets dans un ordre aléatoire.

Au début de ce test, 6 stimuli couvrant l'étendue de la gamme de qualité des phrases synthétisées du test sont présentés aux sujets afin de les familiariser. Pour évaluer la cohérence de chaque sujet, une phrase par configuration est présentée deux fois. Les sujets qui donnent à la même phrase de test deux notes différentes d'au moins deux points sont éliminés. Au total, ce sont donc 60 stimuli (en éliminant les 6 phrases d'apprentissage et la phrase de test de cohérence) qui sont présentés à chaque sujet, totalisant ainsi 1140 notes de qualité.

Lors de ce test, deux sujets parmi les 19 participants ont attribué des notes différant de deux unités aux phrases de contrôle. Ceci nous conduit à l'éviction des résultats de

ces deux sujets. Les 17 sujets qui restent, après un examen préliminaire des notes, semblent avoir adopté des stratégies d'évaluation similaires.

En considérant les MOS moyens, il s'avère que les sujets préfèrent les deux configurations FTR&D et HMM_MFCC_F_0, qui ont eu des MOS pratiquement égaux, par rapport à la configuration HMM_MFCC1. En effet, comme le montre le tableau 4.2 la méthode HMM_MFCC_F_0 obtient un MOS égal à 3.15 contre 3.13 pour la méthode FTR&D et 2.7 pour la méthode HMM_MFCC1. Cette différence significative de MOS entre la configuration HMM_MFCC1 et les deux autres configurations confirme qu'il y a eu une réelle prise en compte du pitch au moment de l'apprentissage et de la classification et que les discontinuités de pitch audible (remarquées dans le test du chapitre précédent) ont bel et bien été réduites, voire éliminées.

Méthode	**MOS**
FTR&D	3.13
HMM_MFCC1	2.77
HMM_MFCC_F_0	3.15

Tableau 4.2 – Résultats des tests MOS

La figure 4.2 présente la différence des moyennes MOS par phrase des deux méthodes HMM_MFCC_$F0$ et FRT&D. Dans cette figure on remarque que pour 11 phrases les sujets ont préféré la méthode HMM_MFCC_F_0 et parmi ces 11 phrases il y a deux phrases pour lesquelles la différence de MOS est presque égale à 1. Pour ces deux phrases nous avons examiné le signal synthétisé par la méthode FTR&D et nous avons remarqué des problèmes de concaténation dus à une discontinuité spectrale.

Deux phrases parmi les 20 phrases ont eu pratiquement les mêmes notes dans les deux configurations. Pour les 7 phrases qui restent les sujets ont préféré la configuration FTR&D. Parmi ces 7 phrases, une seule présente une différence de MOS égal à 1. Après examination du signal de cette phrase synthétisée par la méthode HMM_MFCC_$F0$ il s'est avéré que l'artefact audible est dû à une discontinuité de pitch.

4.2.5.2 Evaluation objective

Pour mesurer les performances de la méthode proposée aux points de concaténations, deux évaluations objectives ont été menées. Pour la première, 1000 phrases du corpus

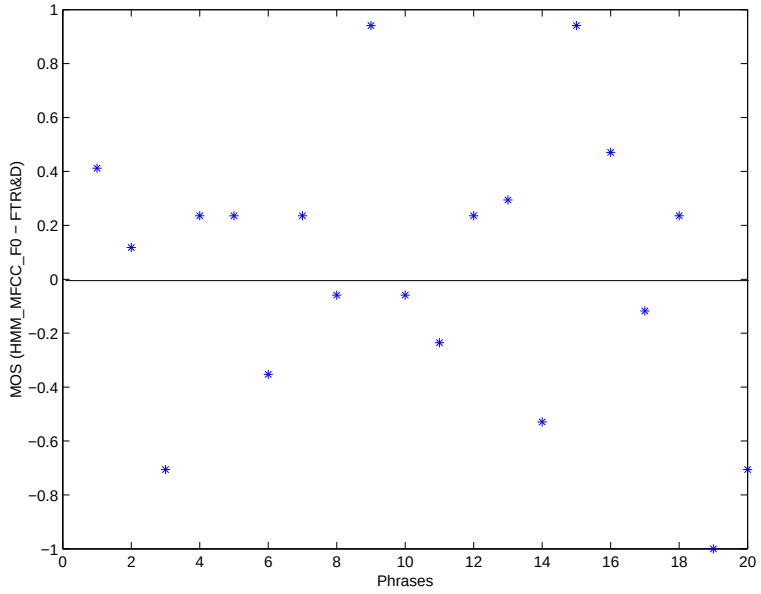


Figure 4.2 – Différences des moyennes MOS par phrase des deux méthodes HMM_MFCC_$F0$ et FRT&D

"Le Monde" ont été synthétisées par les trois configurations, puis une distorsion spectrale normalisée (équation 4.3) et une différence de fréquence fondamentale ont été calculées à chaque point de concaténation. Les histogrammes des distorsions spectrales et des différences de la fréquence fondamentale des trois méthodes sont présentés respectivement dans les figures 4.3 et 4.4.

Méthode	μ_{MFCC}	σ_{MFCC}
FTR&D	4.74	3.38
HMM_MFCC1	3.76	2.66
HMM_MFCC_F_0	3.87	2.69

Tableau 4.3 – Moyennes et écart type de la distorsion spectrale aux points de concaténations pour les trois méthodes testées.

Les histogrammes de distorsion spectrale des deux configurations proposées (HMM_MFCC1, HMM_MFCC_F_0) montrent que les phrases synthétisées par ces deux configurations présentent moins de discontinuités spectrales que celles synthétisées par la configuration FTR&D. Ces résultats sont confirmés par les moyennes et les écarts types du tableau 4.3. Cependant, la petite différence entre les moyennes et les variances des deux

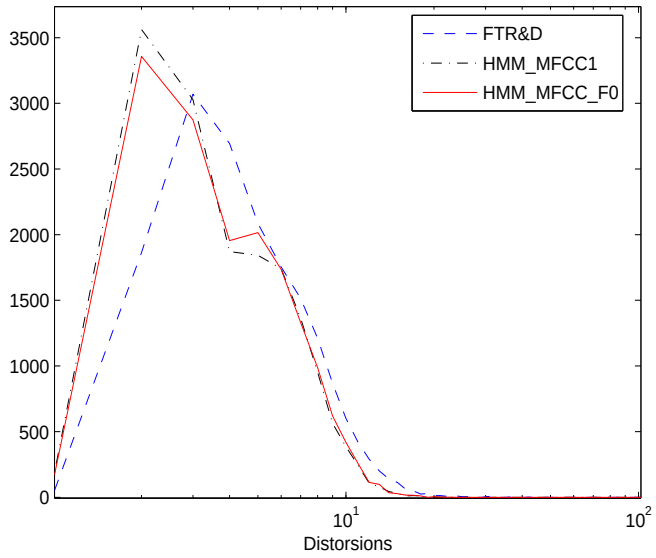


Figure 4.3 – Distorsions spectrales aux points de concaténations calculées pour les trois méthodes

configurations proposées est due à la prise en compte de la fréquence fondamentale dans la configuration HMM_MFCC_F_0. Cela s'explique par le fait que la configuration HMM_MFCC_F_0 perd un peu de continuité spectrale en essayant d'améliorer la continuité de la fréquence fondamentale. Cette constatation est confirmée par les histogrammes de différences de F_0 et leurs moyennes et écarts types présentés dans le tableau 4.4. *À* partir de ces deux derniers (histogrammes de différences de fréquence fondamentale et tableau de moyennes et écarts types) on peut voir clairement l'amélioration apportée par la prise en compte de la fréquence fondamentale au niveau du contrôle de cette dernière aux points de concaténation.

Méthode	μ_{F_0}	σ_{F_0}
FTR&D	5.86	19.86
HMM_MFCC1	7.38	14.81
HMM_MFCC_F_0	4.40	10.50

Tableau 4.4 – Moyennes et écart type de la différence de la fréquence fondamentale aux points de concaténations pour les trois méthodes testées.

L'objectif de la deuxième évaluation est de savoir quel type de concaténation notre

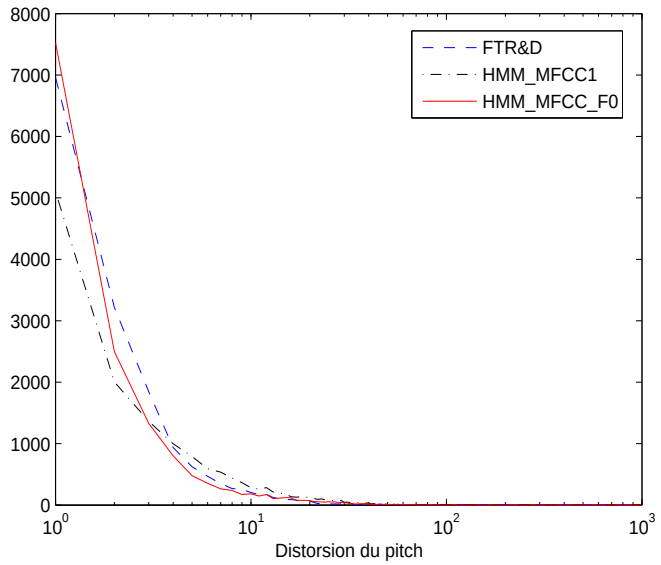


Figure 4.4 – Différences de pitch aux points de concaténations calculées pour les trois méthodes

méthode privilégie. Pour cela, nous avons étudié les concaténations des zones voisées et des zones non voisées séparément. Les distorsions spectrales calculées à partir des phrases synthétisées par la méthode HMM_MFCC_F_0 de la première évaluation ont été utilisées. Les histogrammes des distorsions spectrales des zones voisées et des zones non voisées sont présentés sur la figure 4.5.

La figure 4.5 montre que les distorsions dans les zones voisées sont nettement supérieures aux distorsions dans les zones non voisées. Ces résultats sont confirmés par les moyennes du tableau 4.5. Cela signifie que la méthode proposée privilégie les concaténations sur des zones non voisées ce qui est logique et facile à réaliser, ce qui est un élément en plus pour expliquer les améliorations montrées par les tests effectués auparavant.

Zone de concaténation	μ_{MFCC}
NV	1.7
V	4.76

Tableau 4.5 – Moyennes de la distorsion spectrale aux points de concaténations pour les zones voisées et non voisées.

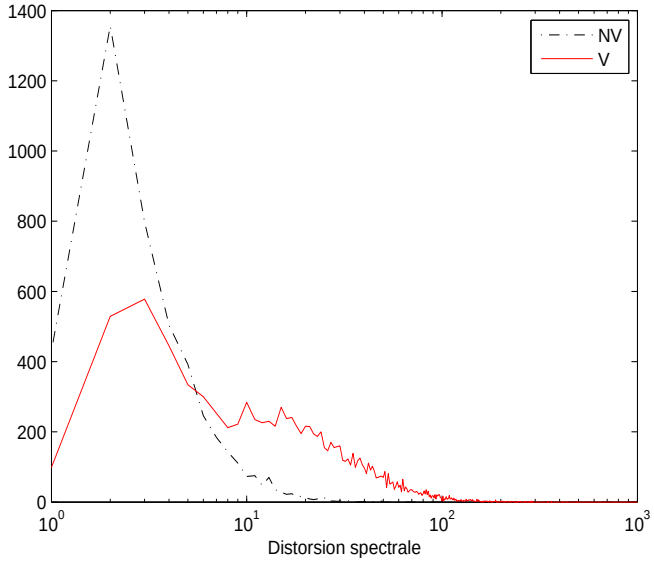


Figure 4.5 – Distorsions spectrales aux points de concaténations calculées pour les zones voisées et non voisées

4.3 Présélection par critères symboliques

Dans la méthode proposée, la sélection des N meilleures instances d'un diphone donné se fait en considérant l'ensemble des représentants de ce diphone présents dans le corpus. Cependant, le nombre de candidats pour un diphone dans la base peut varier (en fonction de la nature du diphone) de quelques représentants à plus de 3000 représentants. Pour réduire la complexité de la sélection des N meilleures instances, une présélection s'avère nécessaire. Ainsi, une présélection purement symbolique a été utilisée.

Dans cette section nous décrivons les différents critères utilisés pour la présélection symbolique, ainsi que la nouvelle architecture du système proposé après l'ajout de la présélection symbolique. Pour clore cette section, nous évaluons par des tests subjectifs et objectifs l'influence de cette présélection symbolique sur la qualité de la parole synthétisée par la méthode proposée.

4.3.1 Critères utilisés pour la présélection

La présélection des unités pour la synthèse d'un groupe de souffle donné est réalisée en fonction des critères symboliques sous forme de règles de filtrage. Ces règles de filtrage permettront d'écarter certains candidats qui sont distants de la meilleure solution. Les différentes règles de filtrages utilisées concernent : les positions syllabiques, les structures syllabiques, les marqueurs mélodiques et le type de mot.

Ces règles sont composées à partir de plusieurs informations, telles que la composition de la syllabe, la position du phone considéré dans la syllabe, la position de la syllabe dans le mot et dans le groupe de souffle, la position du mot dans le groupe de souffle ou encore la position du groupe de souffle dans la phrase. Ces informations sont très riches et permettent de prendre en compte, d'une façon indirecte, des phénomènes acoustiques. Un exemple en est la manière de décrire simplement la nature de la syllabe de façon à refléter la durée et le contour intonatif.

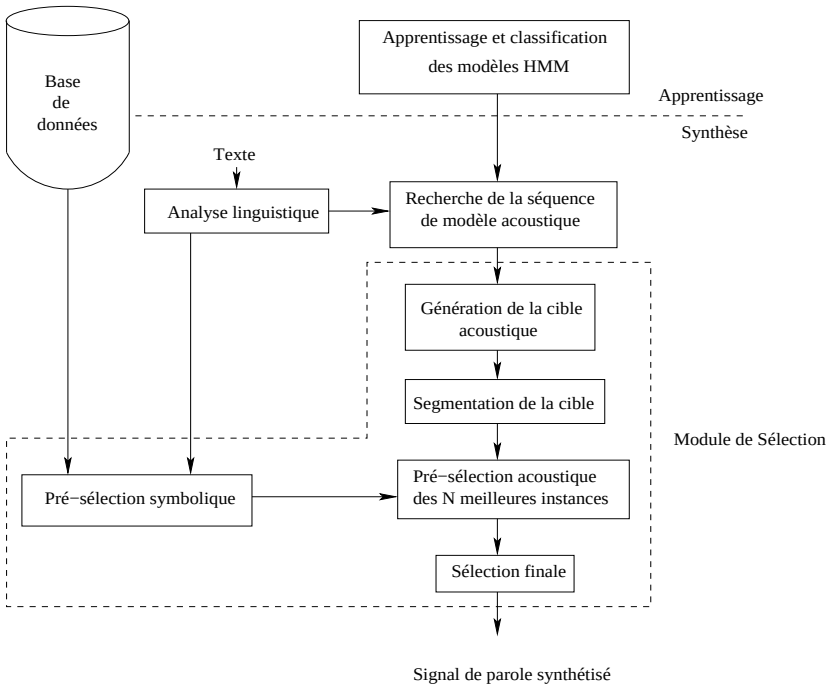


Figure 4.6 – La nouvelle architecture du système de synthèse proposé.

4.3.2 Nouvelle architecture du système proposé

Durant la synthèse, la présélection symbolique est réalisée en même temps que la génération de la cible. Par la suite, les N instances les plus proches de la cible générée sont sélectionnées à partir des candidats déjà présélectionnés symboliquement. Finalement, parmi les N meilleurs instances de tous les diphones, la séquence d'unités qui assure une bonne continuité de la fréquence fondamentale est sélectionnée. La figure 4.6 présente la nouvelle architecture du système proposé avec dans le bloc en pointillés le module de sélection des unités.

4.3.3 Expériences et résultats

Dans cette section, nous présentons les résultats des expérimentations que nous avons effectuées en vue de l'évaluation de l'influence de la pré-sélection symbolique sur la qualité de la parole synthétisée par la méthode proposée.

Dans ces expérimentations, nous avons évalué subjectivement ansi qu'objectivement deux méthodes de synthèse. La première méthode est celle décrite dans ce chapitre (appelée dans cette expérimentation HMM_MFCC_F_0) combinée avec une pré-sélection symbolique. Pour la deuxième méthode, il s'agit de la version actuellement commercialisée du système de synthèse de France Télécom (appelée dans cette expérimentation FTR&D) et qui utilise, elle aussi, la même pré-sélection symbolique. Il est à noter que c'est la version la plus récente du système de synthèse de France Télécom qui est utilisée à chaque test, ce qui explique les petites différences entre les résultats obtenus pour chaque test.

4.3.3.1 Evaluation subjective

Pour évaluer subjectivement l'influence de l'ajout d'une pré-sélection symbolique sur la qualité perçue des phrases synthétisées par la méthode proposée, un test d'écoute a été mené. Ce test d'écoute compare la nouvelle version de la méthode proposée HMM_MFCC_F_0 à la méthode FTR&D. Il est effectué dans les mêmes conditions que celui décrit dans la section précédente et suivant le même protocole (section 4.2.5). Les résultats de ce test sont résumés dans le tableau 4.6. Les auditeurs donnent pratiquement le même score aux deux configurations. En considérant les différences des MOS

par phrase ($MOS_{HMM\_MFCC\_F_0}$ - $MOS_{FTR\&D}$) figure 4.7, pour la moitié des phrases la méthode proposée a été préférée à celle de FTR&D, et une des phrases présente une différence de MOS supérieure à 1. Pour l'autre moitié des phrases, la méthode FTR&D a été préférée par rapport à la méthode proposée avec une différence de MOS qui ne dépasse pas le seuil de 0.7.

Globalement la nouvelle version de la méthode proposée devance celle de FTR&D de quelques centièmes de MOS ce qui n'est pas significatif. Cependant, la complexité du module de sélection de la méthode proposée est significativement améliorée.

Il est à noter que le petit écart entre la méthode proposée et celle de FTR&D est resté pratiquement le même que ce soit avec ou sans l'ajout de la présélection symbolique. Deux éléments permettent d'expliquer la différence de MOS entre les résultats présentés sur les tableaux 4.2 et 4.6. D'une part, la méthode HMM_MFCC utilisée précédemment tend a triée les notes vers le bas. D'autre part, dans le test effectué ici une version plus récente et plus performante du système FTR&D a été utilisée, c'est pourquoi la note MOS obtenue par le système FTR&D est meilleur.

Méthode	**MOS**
FTR&D	3.42
HMM_MFCC_F_0 avec présélection	3.48

Tableau 4.6 – Résultats des tests MOS des deux méthodes FTR&D et HMM_MFCC_F_0

4.3.3.2 Evaluation objective

Pour évaluer objectivement la nouvelle configuration de la méthode proposée (prise en compte du pitch et pré-sélection symbolique), nous nous sommes intéressés à la continuité de la fréquence fondamentale des phrases synthétisées et à leur distorsion spectrale aux points de concaténations. Ce test objectif consiste à mesurer la différence de la fréquence fondamentale et la distorsion spectrale aux points de concaténation pour 1000 phrases du corpus "Le Monde" synthétisées par les deux méthodes (FTR&D, HMM_MFCC_F_0). Les résultats de ce test sont présentés par les histogrammes de la figure 4.8 pour la différence de la fréquence fondamentale et les histogrammes de la figure 4.9 pour la distorsion spectrale.

La figure 4.8 montre que les phrases synthétisées par la méthode proposée présentent

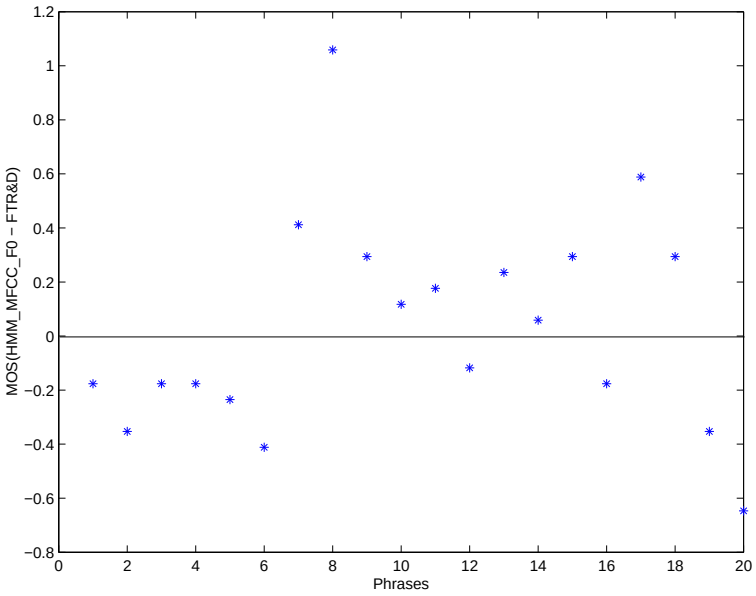


Figure 4.7 – Différences des moyennes MOS par phrase des deux méthodes HMM_MFCC_F_0 et FTR&D

moins de discontinuité de fréquence fondamentale aux points de concaténation que celles synthétisées par la méthode FTR&D. Ces résultats sont confirmés par les moyennes et les écarts types du tableau 4.7. Cela s'explique par le fait qu'il y a déjà une prise en compte du pitch au moment de la pré-sélection symbolique et qui est en plus raffinée par la suite par l'utilisation de la cible acoustique.

Méthode	μ_{F_0}	σ_{F_0}
FTR&D	5.86	19.86
HMM_MFCC_F_0 avec présélection	1.91	7.15
HMM_MFCC_F_0	4.40	10.50

Tableau 4.7 – Moyennes et écarts types de la différence du pitch aux points de concaténation pour les deux méthodes FTR&D et HMM_MFCC_F_0 avec et sans présélection

Au niveau de la distorsion spectrale, les deux méthodes testées présentent des résultats similaires comme le montrent la figure 4.9 et le tableau 4.8. Cependant, la configuration HMM_MFCC_F_0 comparée à la configuration présentée dans la section précédente (sans la pré-sélection symbolique) présente une légère perte du contrôle de

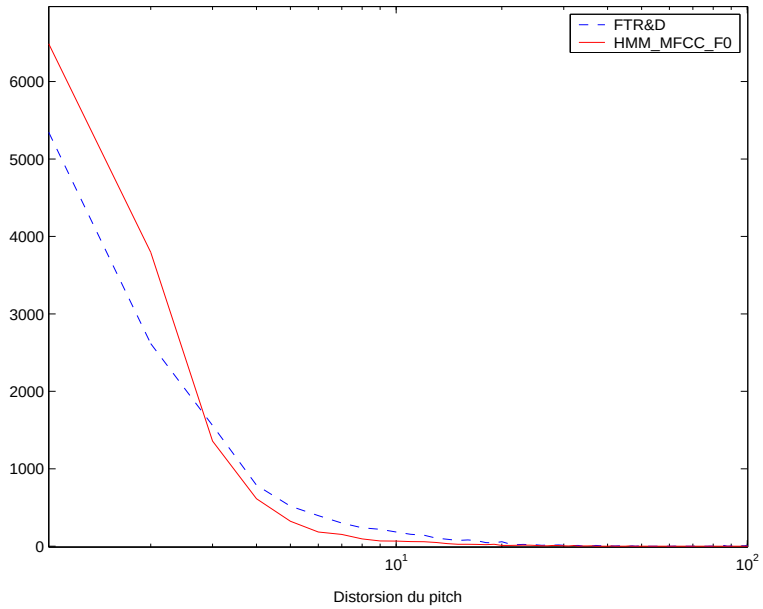


Figure 4.8 – Différences de fréquence fondamentale calculées aux points de concaténation pour les deux méthodes FTR&D et HMM_MFCC_F_0 avec présélection.

la distorsion spectrale aux points de concaténations. Cette perte est due au nombre restreint d'instances en sortie de la pré-sélection symbolique, ces instances étant par la suite utilisées pour la pré-sélection acoustique. En revanche, cette perte est négligeable par rapport à la bonne maîtrise de la continuité du pitch et la réduction significative de la complexité du module de sélection.

Méthode	μ_{MFCC}	σ_{MFCC}
FTR&D	4.74	3.38
HMM_MFCC_F_0 avec présélection	4.95	3.38
HMM_MFCC_F_0	3.87	2.69

Tableau 4.8 – Moyennes et écarts types de la distorsion spectrale aux points de concaténations pour les deux méthodes FTR&D et HMM_MFCC_F_0 avec et sans présélection

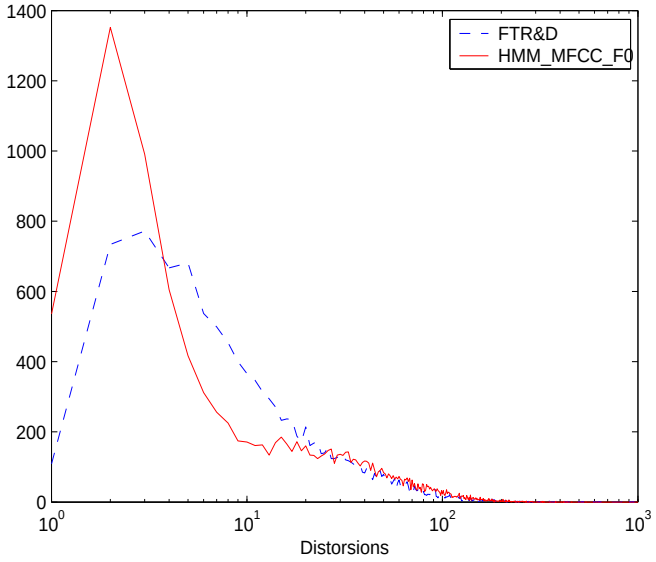


Figure 4.9 – Distorsions spectrales calculées aux points de concaténation pour les deux méthodes FTR&D et HMM_MFCC_F_0 avec présélection.

4.4 Conclusion

Dans ce chapitre, nous avons apporté des modifications à la méthode de sélection proposée au troisième chapitre en vue de pallier certaines déficiences. Tout d'abord, nous avons proposé une stratégie de sélection ayant un meilleur contrôle de la fréquence fondamentale. Pour cela, nous avons intégré la fréquence fondamentale dans les vecteurs acoustiques de façon à tenir explicitement compte de cette information lors de la phase de classification. Le processus de sélection lui-même a fait l'objet de modification ; d'une part la cible acoustique est uniquement utilisée à des fins de présélection et d'autre part un coût de concaténation basé sur une différence de fréquence fondamentale est introduit de manière à minimiser les discontinuités de la fréquence fondamentale de la phrase synthétisée.

Des tests tant subjectifs qu'objectifs ont été menés pour évaluer cette prise en compte de la fréquence fondamentale. Les résultats de ces tests ont montré une nette amélioration de la continuité de la fréquence fondamentale de la version présentée dans ce chapitre par rapport à celle proposée au chapitre précédent. De plus, les phrases synthétisées par la nouvelle version de la méthode proposée présentent une

meilleure continuité spectrale comparées aux même phrases synthétisées par le système de synthèse de France Télécom.

Par ailleurs, en vue de réduire la complexité de notre méthode, nous avons utilisé une présélection symbolique. Cette présélection consiste à éliminer, sur la base d'informations symboliques extraites du texte à synthétiser, les unités les moins appropriées au contexte de synthèse.

Des tests subjectifs et objectifs ont été menés pour évaluer l'influence de l'ajout de la présélection symbolique sur la qualité de la parole synthétisée par la méthode proposée. Ils montrent que la qualité est restée pratiquement la même, que ce soit au niveau de la qualité perçue ou de la continuité de la fréquence fondamentale. Cependant, la complexité de la méthode proposée est significativement réduite.

Chapitre 5

Réduction de bases à partir de critères acoustiques

5.1 Introduction

Dans l'état de l'art, la majorité des systèmes TTS existants sont basés sur une synthèse par corpus. Dans ces systèmes, les meilleures instances acoustiques sont sélectionnées dans une grande base d'unités acoustiques, puis concaténées pour générer de la parole synthétique. Pour ce type de système, la synthèse de la parole devient un problème de collection, d'annotation et de recherche dans un corpus de parole pré-enregistré [HB96] [HAA$^+$96] [CYC01]. La qualité de la parole synthétisée dépend dans une large mesure de la taille et surtout de la couverture acoustique du corpus. Cependant, l'utilisation de corpus de taille importante (plusieurs centaines de Mo, correspondant à plusieurs heures de parole enregistrée) n'est pas sans poser de problème.

En effet, l'algorithme de Viterbi utilisé pour la sélection des unités acoustiques a une complexité en $O(M^2N)$ où N est le nombre d'unités de la phrase à synthétiser et M le nombre moyen d'unités acoustiques pour chaque unité de synthèse. Pour réduire cette complexité une réduction de la largeur du treillis est donc souhaitable. L'utilisation de bases acoustiques de taille réduite est également nécessaire pour l'intégration des systèmes TTS dans des terminaux dont les capacités en terme de mémoire mais aussi de CPU peuvent être limitées.

Il est donc crucial, lors de la conception d'un système de synthèse par corpus, d'opti-

miser la base acoustique effectivement utilisée. Cette optimisation consiste à minimiser la taille d'un dictionnaire acoustique tout en préservant une couverture acoustique acceptable. Dans ce cadre, il s'agit, d'une part, d'éliminer les unités acoustiques redondantes, d'autre part, d'écarter les unités acoustiques dont les caractéristiques sont telles qu'elles seraient difficilement utilisables à la synthèse. Plus précisément, des défauts d'articulations peuvent subvenir lors de l'enregistrement du corpus, ce qui conduit par exemple à l'apparition d'unités acoustiques trop courtes ou trop longues. En outre, certaines peuvent avoir été prononcées dans des configurations prosodiques extrêmes. En particulier, des unités acoustiques ayant des valeurs de fréquence fondamentale trop élevées ou trop faibles peuvent être présentes dans le corpus, mais seraient difficilement utilisables en synthèse, car incompatibles avec la majorité des autres unités acoustiques. Enfin, des phénomènes de coarticulation très intenses peuvent apparaître rendant la segmentation des unités acoustiques très délicate. De telles unités acoustiques seront donc difficiles à classifier et donc à maîtriser sur le plan acoustique.

Dans ce chapitre nous nous intéressons à l'application de la méthode proposée aux chapitres 3 et 4 dans le cadre de la réduction de bases. La première partie de ce chapitre est dédiée aux différentes méthodes de réduction de bases existant dans la littérature. La deuxième partie concerne la méthode proposée et sa mise en œuvre pour la réduction de bases. Enfin, dans une dernière partie nous présentons les résultats des expérimentations que nous avons effectuées pour évaluer la méthode proposée.

5.2 État de l'art des méthodes de réduction de bases

Dans les systèmes de synthèse par corpus actuels, il y a principalement deux approches pour la réduction de bases. La première, dite réduction *a priori* [BT97] [Don01] [YG02], est basée sur une classification des instances d'unités en fonction des informations contextuelles et acoustiques de chaque unité acoustique. La seconde, dite réduction *a posteriori* [Don00], [CBSB00], [RAFT02] est basée sur l'analyse des fréquences d'utilisation des différentes unités acoustiques lors de la synthèse, l'objectif étant dans ce cas de conserver les unités acoustiques sélectionnées le plus fréquemment.

5.2.1 Les méthodes de réduction *a priori*

La réduction de base *a priori* vise à rechercher les unités acoustiques les plus pertinentes à partir de descripteurs symboliques voire acoustiques. L'objectif est d'une part de limiter la redondance au sein du corpus et d'autre part de supprimer des unités apparaissant comme singulières, c'est-à-dire dont les caractéristiques acoustiques sont telles qu'elles seraient difficilement compatibles avec d'autres unités acoustiques. En général, de telles unités résultent de défauts de prononciation voire d'erreurs dans la segmentation phonétique. Ce type de réduction peut être obtenu après classification explicite des unités acoustiques. Des techniques de classification hiérarchique descendante à base d'arbres de décision peuvent par exemple être utilisées. Dans [BT97], un tel arbre de décision permet tout d'abord de classer les unités en fonction de leurs contextes phonétique et prosodique, sur la base de critères acoustiques (e.g. variance des MFCC, de l'énergie ou de la fréquence fondamentale). Ensuite, pour chaque feuille de l'arbre, une unité moyenne ou centroïde est déterminée et les unités les plus éloignées de ce centroïde sont éliminées. Une classification plus fine est également réalisée au sein de chaque feuille de manière à réduire la redondance acoustique. Ce type de technique a également été repris dans [LHSW04] où une distinction entre unités voisées et non voisées est faite. Pour les unités voisées, la classification prend en compte l'information d'enveloppe spectrale et la fréquence fondamentale, alors que pour les unités non voisées, seule l'enveloppe spectrale est utilisée. Par ailleurs, d'autres méthodes ne reposant pas sur une classification explicite des unités acoustiques ont été proposées. Ces méthodes permettent en général de localiser les unités singulières. Le principe est de définir un prototype de chaque unité symbolique, puis d'effectuer un classement des unités acoustiques correspondantes en fonction de leur distance à ce prototype. Cette méthode a été appliquée dans [YWZ+04] dans un contexte de synthèse en langue chinoise par concaténation de syllabes. Pour cela, différentes informations (durée, énergie, fréquence fondamentale et MFCC) ont été prises en compte, tout d'abord séparément, puis de manière combinée. Dans [KK04], une mesure de compatibilité prosodique est définie pour chaque unité syllabique. Celle-ci permet, dans un premier temps, d'éliminer les unités acoustiques qui donneraient lieu en moyenne à de mauvaises concaténations. A partir de cet ensemble initial et selon la taille de la base souhaitée, d'autres unités acoustiques sont ajoutées. L'ajout de ces unités est réalisé de manière itérative en minimisant une fonction de coût composite incluant la mesure de compatibilité précédemment

définie et un terme de répulsion, permettant d'augmenter la diversité prosodique de la base résultante.

5.2.2 Méthodes de réduction *a posteriori*

Ces méthodes de réduction de bases dépendent essentiellement de la méthode de sélection utilisée par le synthétiseur. Ainsi Black dans [BC95] synthétise un grand nombre de phrases, pour ne garder ensuite que les unités acoustiques qui ont les fréquences d'occurrence les plus élevées. Outre cette information, Ling et Zhang prennent en compte dans [LHSW04] une mesure de concaténation des unités acoustiques effectivement utilisées à la synthèse. Ce faisant, ils éliminent celles dont le coût de concaténation moyen est trop élevé.

La difficulté principale de ce type d'approche basée sur la fréquence d'occurrence est d'une part, qu'elle ne fait pas la différence entre les unités acoustiques similaires, autrement dit elle tend à conserver toutes les unités dont les fréquences d'occurrence sont élevées même si elles se ressemblent. De ce fait, ce genre de technique ne permet pas de contrôler la redondance de la base réduite. D'autre part, un simple critère de réduction basé sur la fréquence d'occurrence risque d'occasionner la suppression des unités les plus rares. Le fait de plus garantir la couverture d'un ensemble minimal d'unités fait que le système de synthèse utilisant un tel corpus réduit ne serait plus capable de vocaliser n'importe quel texte. Pour remédier à ces problèmes, Conkie propose dans [CBSB00] de former une base à partir des séquences de triphones sélectionnées au cours de la synthèse par demi-phones d'un grand nombre de phrases. Si une séquence de triphones particulière n'est pas présente dans la nouvelle base réduite, le système la cherche dans la base d'origine. Similairement, Rutten et al. dans [RAFT02] conservent les unités acoustiques sélectionnées en prenant en compte deux critères : la fréquence d'occurrence la plus élevée et la rareté dans le corpus.

D'autres auteurs ont voulu profiter des avantages des deux méthodes en utilisant la fréquence d'occurrence comme information pour une éventuelle classification. Ainsi, Sanghun et al. proposent dans [SYK01] une méthode de réduction basée sur une quantification vectorielle pondérée. Avant la classification des unités acoustiques, les auteurs procèdent à un comptage d'occurrences des instances en synthétisant un ensemble de phrases. Ce taux d'apparition est utilisé, au moment de la classification, comme un fac-

teur de pondération pour les différents paramètres prosodiques du vecteur de quantification. Dans [ZCPC03] la quantification vectorielle pondérée est réutilisée mais associée à d'autres critères de réduction. Le premier d'entre eux repose sur la notion d'importance de l'unité acoustique dans la base, définie comme la fréquence de sélection de cette unité acoustique divisée par la somme des fréquences de sélection de toutes les unités acoustiques correspondant à cette même unité symbolique. Le deuxième concerne la variabilité prosodique de l'unité acoustique. Ces deux critères sont testés séparément puis combinés.

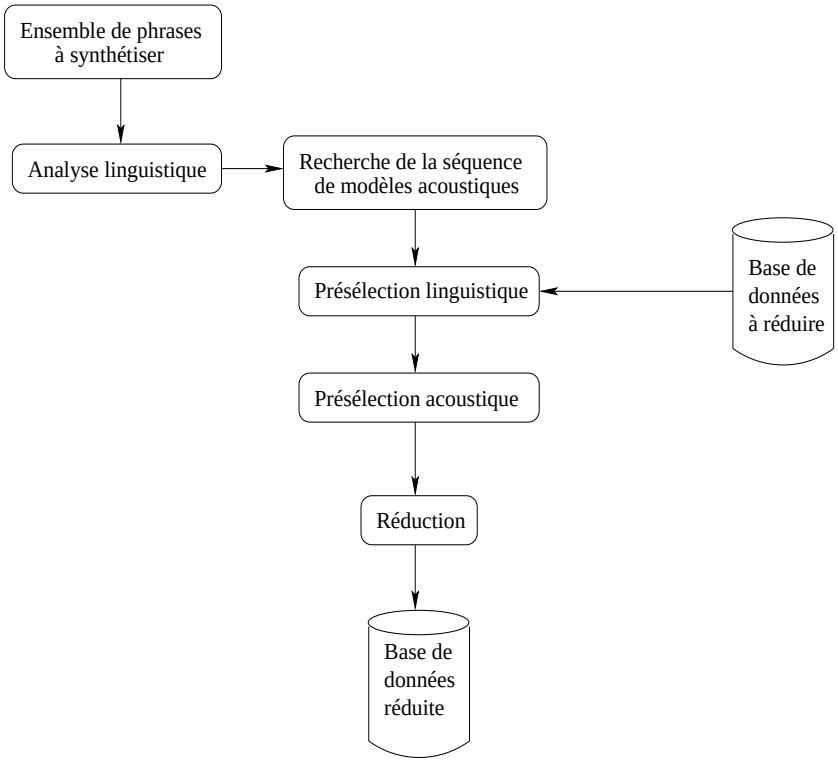


Figure 5.1 – Différentes étapes de la réduction de bases par la méthode proposée.

5.3 Méthode de réduction proposée

5.3.1 Principe

Dans cette section, nous proposons une nouvelle méthode de réduction basée sur la technique de sélection présentée aux chapitres 3 et 4. La méthode proposée s'inscrit

dans le cadre d'une réduction de bases *a posteriori*. La figure 5.1 présente les différentes étapes de la méthode de réduction proposée. De manière schématique, il s'agit de mettre en œuvre sur un corpus textuel de grande taille les modules de présélection décrit au chapitre 4 pour sélectionner un ensemble d'unités candidates. Une analyse des statistiques d'utilisation des unités ainsi présélectionnées permet ensuite de déterminer les unités les plus pertinentes pour la synthèse. Le traitement commence par l'analyse linguistique du texte en entrée du système. A partir de la description symbolique fournie par les étages linguistiques, une présélection symbolique est effectuée par le module présenté en section 4.3.1. Le module de présélection acoustique décrit en 4.2.4 délivre, pour chaque unité de synthèse, un nombre d'unités candidates au plus égal à N. Lorsqu'une unité acoustique est ainsi présélectionnée, le nombre d'utilisation de cette unité est incrémenté de 1. L'application de cette séquence de traitement sur un ensemble de M phrases conduit à l'obtention d'un histogramme d'utilisation des différentes unités du corpus. La réduction consiste alors à décider, sur la base de cet histogramme, voire sur la base de règles complémentaires, des unités qu'il convient de conserver dans le corpus de synthèse. Bien entendu, pour que ce type de méthode *a posteriori* soit efficace, il faut disposer d'un nombre suffisant de phrases afin de détecter de manière fiable les unités les plus pertinentes. Il convient également bien entendu de disposer d'un ensemble de phrases suffisamment riche pour couvrir le maximum de contextes de synthèse possibles. Notons de plus qu'une méthode de réduction de dictionnaire acoustique basée sur des considérations de fréquences d'occurrences des unités acoustiques n'est justifiable que si l'ensemble de textes à synthétiser est tel que la distribution des unités symboliques qui y sont contenues suit la distribution des unités symboliques du corpus de synthèse original. La détermination d'un tel corpus textuel est délicate et sort du contexte de cette thèse. Néanmoins, nous pouvons constater qu'un corpus de synthèse contient généralement un nombre relativement restreint d'unités fortement représentées et un nombre important d'unités rares. Il est clair que ces unités rares sont à conserver, sans quoi, le système de synthèse ne serait plus à même de vocaliser une phrase contenant au moins une telle unité. En revanche, lorsqu'une unité symbolique est fortement représentée, il s'agit d'éliminer les unités acoustiques qui sont peu ou pas choisies. La méthode proposée s'intéresse donc uniquement à l'éviction de ces unités acoustiques, et ceci dans le cas où l'unité de synthèse est le diphone.

5.3.2 Mise en œuvre

Pour mettre en œuvre notre procédure de réduction, nous avons utilisé un ensemble de 20000 phrases extraites du corpus Le Monde. Le corpus de synthèse original est le corpus de la voix Philippe. Les traitements de la figure 5.1 ont été menés en considérant les valeurs suivantes du nombre maximal N d'unités candidates : 5, 10, et 20. A l'issue de ces traitements, toutes les unités ayant été présélectionnées au moins une fois sont conservées. De plus, pour veiller à conserver les unités rares, nous effectuons un post-traitement qui consiste à réinjecter l'ensemble des instances des diphones rares, un diphone étant considéré comme rare s'il contient moins de 5 représentants acoustiques. Une fois ces traitements effectués, nous vérifions bien que chaque diphone ne faisant pas partie des diphones rares dispose d'au moins 5 représentants dans le corpus réduit. Le tableau 5.1 donne les tailles des bases réduites relativement au corpus original pour les différentes valeurs de N testées. Nous remarquons donc, comme cela était prévisible, que plus le paramètre N est élevé plus la base résultante se trouve réduite.

5.4 Expérimentation et résultats

Dans cette section, nous présentons les résultats des expérimentations que nous avons effectuées en vue de l'évaluation de la méthode de réduction de bases proposée avec différents taux de réduction.

Dans ces expérimentations, nous avons réduit la base Philippe (appelée dans cette expérimentation TB) en synthétisant un corpus de 20000 phrases et en testant différentes valeurs de N meilleures instances (5, 10 et 20). Le taux de réduction pour chaque valeur de N testée et l'appellation de chaque base réduite résultante sont résumés dans le tableau 5.1.

N	**Taille relative de la base**	**Appellation**
5	29.65%	BD3
10	45.23%	BD2
20	59.75%	BD1

Tableau 5.1 – Les différentes bases réduites testées dans cette expérimentation

5.4.1 Evaluation subjective

Pour évaluer subjectivement la méthode de réduction de bases proposée, un test de qualité globale a été mené. Ce dernier est effectué dans les mêmes conditions que les tests d'écoute décrits dans le chapitre précédent et suivant les mêmes recommandations (4.2.5.1). Les résultats de ce test sont présentés dans le tableau 5.2. Les auditeurs donnent presque le même score aux quatre bases avec une légère préférence aux phrases synthétisées à partir de la base BD2.

Base	**MOS**
TB	3.22
BD1	3.21
BD2	3.31
BD3	3.26

Tableau 5.2 – Résultats des tests MOS de la méthode proposée avec les différentes bases

La figure 5.2 présente pour chacune des phrases testées, les différences de MOS obtenues en considérant d'une part la base complète (TB) et les différentes bases réduites (BD1, BD2, BD3). Considérant les différences des MOS entre BD2 et TB, pour 10 phrases la synthèse à partir de la BD2 a été préférée à celles synthétisées à partir de TB. Pour 4 phrases, les deux bases ont eu le même score et pour les 6 autres phrases la synthèse à partir de la base TB a été préférée à celles obtenues à partir de la base BD2. Dans le même type de comparaison mais avec d'autres bases (BD3 vs TB), pour 11 phrases la synthèse à partir de la BD3 a été préférée à celles synthétisées à partir de TB. Pour les 9 autres phrases la synthèse à partir de la bases TB a été préférée de celle de la base BD3. La troisième comparaison est entre BD1 et TB. Pour 10 phrases la synthèse à partir de la BD1 a été préférée à celles synthétisées à partir de TB. Pour 1 phrase les deux bases ont eu le même score et pour les 9 autres phrases la synthèse à partir de la bases TB a été préférée de celle de la base BD1.

Globalement la qualité subjective des phrases synthétisées reste la même si la base est réduite à 59.75%, elle est un peu meilleure à 45.23% et d'une façon moins significative à 29.65%. On arrive à réduire fortement la taille de la base tout en gardant une bonne qualité globale. Cela s'explique par le fait que les unités acoustiques éliminées ne servent pas à la synthèse, soit parce qu'elles représentent des unités acoustiques redondantes

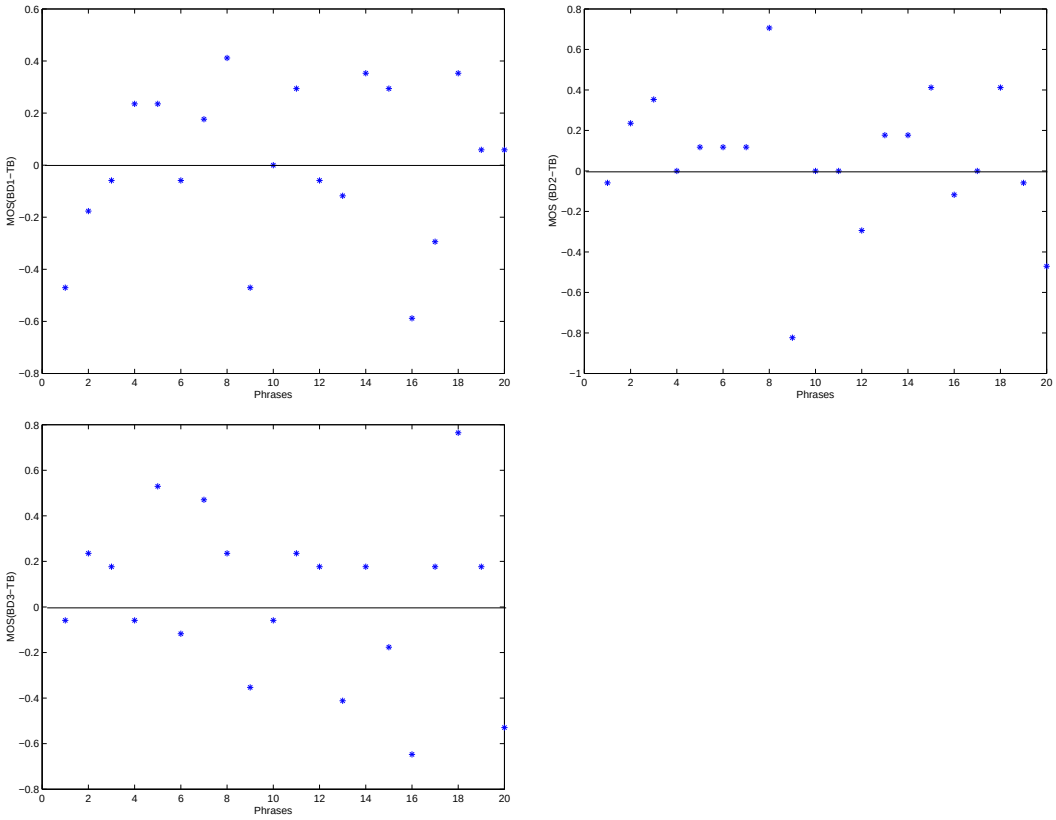


Figure 5.2 – Différences des MOS par phrase de chacune des bases réduites comparées à la base entière.

ou à cause des possibles problèmes d'articulations ou de segmentations.

En d'autres termes, l'effet positif de notre procédure de réduction de base est que nous avons réduit la variabilité acoustique, en éliminant des unités acoustiques trop singulières et qui auraient pour effet d'introduire des artefacts à la synthèse. Cependant, une réduction trop importante conduit à l'apparition de "trous" de couverture acoustique préjudiciables à la qualité de la parole synthétique produite. Ce phénomène est d'ailleurs généralement observé dans les études consacrées à la réduction de base acoustique [RAFT02]. En revanche, la détermination automatique du taux de compression conduisant à une qualité optimale reste un problème ouvert.

5.4.2 Evaluation objective

Pour mesurer les performances de la méthode proposée sur le plan objectif, nous avons utilisé la distortion spectrale (4.3) et la différence de la fréquence fondamentale décrites dans le chapitre précédent. Ces deux distorsions ont été calculés à chaque point de concaténation après la synthèse de 1000 phrases du corpus "Le Monde".

La figure 5.3 illustre les histogrammes de la distorsion spectrale de toutes les bases testées. En effet, les histogrammes de la figure 5.3 montrent que les phrases synthétisées à partir des trois bases réduites présentent moins de discontinuité spectrale aux points de concaténation que celles synthétisées à partir de la base entière. Ces résultats sont confirmés par les moyennes et les écarts types du tableau 5.3. Par ailleurs, la différence entre les trois bases réduites est négligeable au niveau de la distorsion spectrale. D'une part, cela signifie qu'une bonne partie des instances défectueuses ont été éliminées en réduisant la base à 59.75%, d'autre part, les similitudes des résultats entre les trois bases réduites s'expliquent par le fait que les instances éliminées ressemblent acoustiquement à celles qui ont été conservées.

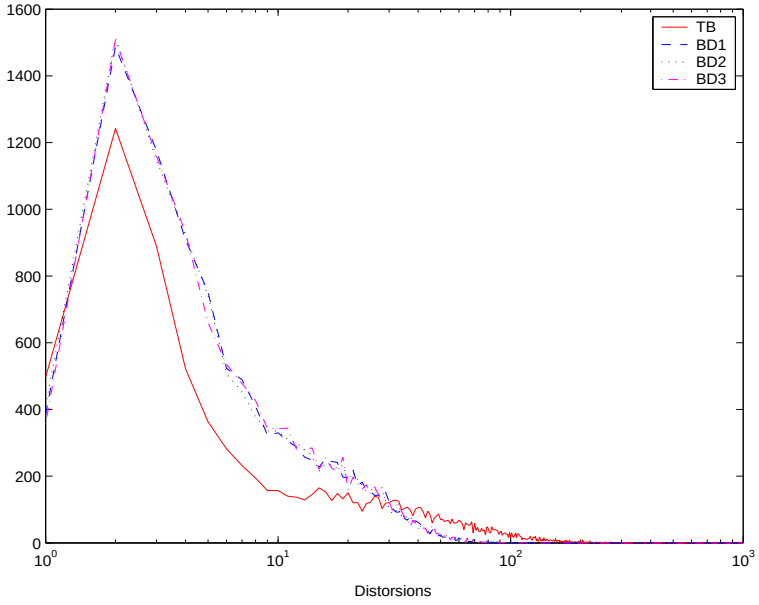


Figure 5.3 – Distorsion spectrale calculée aux points de concaténations pour la méthode proposée avec les différentes bases.

Sur le plan de la différence de la fréquence fondamentale figure 5.4, les phrases

Base	μ_{MFCC}	σ_{MFCC}
TB	4.91	3.36
BD3	3.09	1.78
BD2	3.08	1.8
BD1	3.08	1.78

Tableau 5.3 – Moyennes et écarts types de la distorsion spectrale aux points de concaténations de la méthode proposée avec les différentes bases.

synthétisées à partir des différentes bases testées présentent, globalement, la même discontinuité de fréquence fondamentale aux points de concaténation. Cela est confirmé par les moyennes et les écarts types du tableau 5.4. Ces résultats confirment les observations faites au paragraphe précédent (i.e. la réduction de la base maintient voire améliore la qualité de la synthèse).

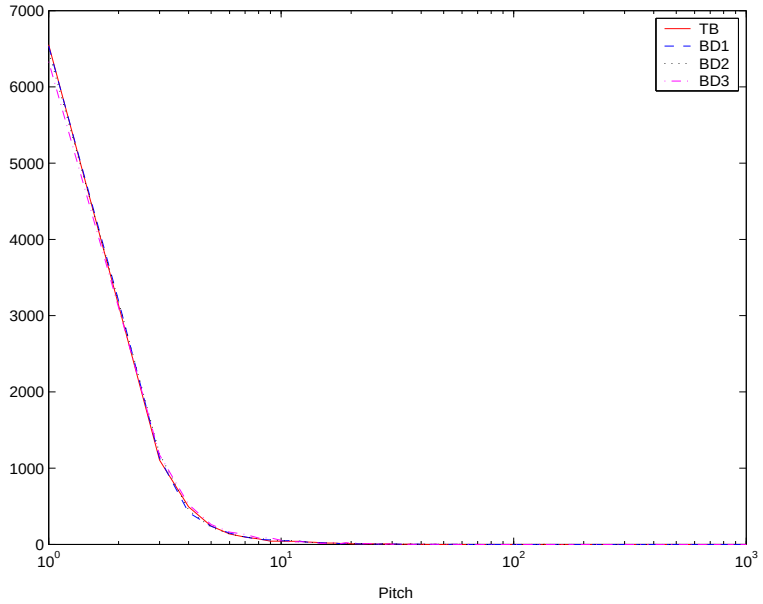


Figure 5.4 – Différences de la fréquence fondamentale calculées aux points de concaténations pour la méthode proposée avec les différentes bases.

Base	μ_{F_0}	σ_{F_0}
TB	1.91	7.15
BD3	1.48	3.85
BD2	1.36	3.59
BD1	1.32	3.46

Tableau 5.4 – Moyennes et écarts types de la différence de la fréquence fondamentale aux points de concaténations de la méthode proposée avec les différentes bases

5.5 Conclusion

Dans ce chapitre nous avons proposé une méthode de réduction de bases s'appuyant essentiellement sur des critères acoustiques. La méthode proposée opère en deux étapes.

Durant l'étape de préparation des données, effectuée hors ligne, un apprentissage des modèles acoustiques est réalisé sur les données de la base à réduire. Ensuite, une classification de ces modèles par arbre de décision est menée. L'étape suivante consiste à synthétiser un ensemble de phrases. Durant la synthèse, une présélection symbolique en même temps qu'une génération d'une cible acoustique sont réalisées pour chaque phrase de l'ensemble. Par la suite, les N meilleures unités acoustiques les plus proches de la cible générée sont sélectionnées. Ces dernières sont stockées pour constituer la nouvelle base réduite.

Plusieurs taux de réductions ont été testés subjectivement ainsi qu'objectivement. Les résultats de ces tests montrent qu'une réduction à 45.23% améliore la qualité de la synthèse sur les deux plans (subjectif et objectif). En revanche, pour une réduction à 29.65% l'amélioration n'est plus significative. En gardant 59.75% de la base, la qualité de synthèse est pratiquement la même.

Des expérimentations supplémentaires mériteraient d'être menées afin d'explorer les limites de la méthode proposée, par exemple, utiliser une base de test beaucoup plus grande et chercher le taux de réduction le plus faible après lequel la qualité de synthèse se dégrade significativement.

Conclusion

Le travail réalisé au cours de cette thèse a porté sur la recherche d'une nouvelle procédure de sélection des unités acoustiques, pour la synthèse vocale par corpus. L'objectif de cette étude est double : d'une part d'introduire des critères acoustiques dans le processus de sélection des unités afin de limiter les discontinuités acoustiques et d'autre part d'utiliser cette nouvelle procédure dans un but de réduction des bases acoustiques.

Contexte du travail et problématique

Ce travail s'inscrit dans le cadre de la synthèse par corpus. Les performances de tels systèmes de synthèse reposent sur deux facteurs, à savoir d'une part la représentativité du corpus de synthèse et d'autre part l'algorithmie de sélection capable d'exploiter toute cette richesse. Malgré les bons résultats fournis par cette technique, la synthèse par corpus se heurte à un problème de robustesse, c'est-à-dire qu'elle n'est pas capable de garantir une parole de très haute qualité sur l'ensemble d'un énoncé. Les déficiences des systèmes de SPC sont dans une très large mesure dues à la méthode de sélection utilisée. En effet, les techniques de sélection actuelles font essentiellement intervenir des coûts cible symboliques et de ce fait il est très difficile de pouvoir maîtriser sur le plan acoustique le signal synthétiser. L'objectif de cette thèse est donc d'introduire des cibles acoustiques dans le processus de sélection des unités.

La première partie de ce document s'est intéressée au cadre général de nos travaux. Après avoir abordé la synthèse de la parole, et présenté les différentes étapes du processus de cette synthèse à partir de la représentation textuelle et les différents méthodes de génération sonore, nous avons dressé un inventaire des différentes unités utilisées en synthèse par concaténation. Ensuite, nous avons introduit la synthèse par

corpus en deux parties. Dans la première partie, nous avons présenté les principes de la sélection des unités ainsi qu'un état de l'art des méthodes manuelles, de la sélection des unités acoustiques ainsi que les avantages et les inconvénients de ces méthodes. Dans la deuxième partie, nous avons abordé l'automatisation de la procédure de sélection et dressé un inventaire des différents systèmes utilisant ces techniques de sélection. À la fin de cette troisième partie, nous avons tiré les enseignements de ces études bibliographiques et souligné les lacunes des systèmes de synthèse actuels.

Contributions et résultats

La contribution principale de cette thèse est la mise en œuvre d'une nouvelle approche de sélection des unités basée sur la génération explicite d'une cible acoustique. La méthode proposée, détaillée au chapitre 3, consiste dans un premier temps à effectuer un apprentissage de modèles HMM puis une classification acoustique des différents modèles de sénones résultants via des arbres de décision. Ces modèles de sénones sont ensuite utilisés dans la phase de synthèse pour définir une cible acoustique, laquelle sert de coût cible dans le processus de sélection.

L'évaluation de notre approche a été effectuée, sous la forme de tests d'écoute formels, en comparant un ensemble de phrases synthétisées par notre approche à la synthèse du même ensemble de phrases par la procédure actuellement utilisée dans le système de synthèse de FTR&D. Un point intéressant est que la méthode proposée conduit à une meilleure continuité spectrale des unités sélectionnées. Néanmoins, des discontinuités de la fréquence fondamentale ont été remarquées, de sorte qu'au final les tests font apparaître une légère préférence pour la procédure de sélection du système de synthèse de FTR&D. Ceci étant cette étude a permis de mettre en évidence toute la flexibilité de notre approche. D'une part, la méthode de sélection proposée est complètement automatisée et donc directement transposable à d'autres voix, voire à d'autres langues. D'autre part, bien qu'ayant été présentée dans le cadre de la sélection de diphones, la méthodologie peut être adaptée très facilement à la sélection de tout autre type d'unités (demi-phones, sénones, ...).

Sur la base de ces résultats nous avons apporté, dans le chapitre 4, des modifications de la méthode de sélection proposée. Tout d'abord, nous avons proposé une stratégie de sélection ayant un meilleur contrôle de la fréquence fondamentale. Pour cela, nous

avons intégré la fréquence fondamentale dans les vecteurs acoustiques de façon à tenir explicitement compte de cette information lors de la phase de classification. Le processus de sélection lui-même a fait l'objet de modification ; d'une part la cible acoustique est uniquement utilisée à des fins de présélection et d'autre part un coût de concaténation basé sur une différence de fréquence fondamentale est introduit de manière à minimiser les discontinuités de la fréquence fondamentale de la phrase synthétisée.

Par ailleurs, en vue de réduire la complexité de notre méthode, nous avons effectué une présélection symbolique en amont de la présélection acoustique. Cette présélection a pour effet d'éliminer, sur la base d'informations symboliques extraites du texte à synthétiser, les unités les moins appropriées au contexte de synthèse. La nouvelle version de la méthode proposée a été testée tant sur les plans subjectif qu'objectif. Les résultats des tests ont montré une nette amélioration de la continuité de la fréquence fondamentale. En outre, l'ajout de la présélection symbolique a permis de réduire significativement la complexité tout en gardant un bon niveau de qualité perçue.

Dans le chapitre 5, nous avons utilisé le formalisme développé dans les deux derniers chapitres pour concevoir un algorithme de réduction de bases à partir de critères acoustiques. La réduction est réalisée en gardant l'ensemble des unités acoustiques résultant de la synthèse d'un corpus textuel de grande taille. Plusieurs taux de réduction ont été testés subjectivement ainsi qu'objectivement. Les résultats des tests ont montré qu'on peut éliminer jusqu'à 60% de la base tout en conservant une bonne qualité globale.

Perspectives

Nous avons mis en œuvre une procédure de sélection des unités acoustiques basée essentiellement sur des critères acoustiques. L'évaluation de notre travail a été menée sur un seul corpus de synthèse. Aussi, pour valider la méthode proposée, il serait intéressant de la tester sur d'autres voix, voire sur d'autres langues.

Il convient également de comparer la technique de sélection proposée qui sélectionne les unités candidates les plus proches d'une cible acoustique via un algorithme de DTW à des méthodes basées sur la maximisation de la vraisemblance pour la sélection de la séquence optimale [Don96].

Pour améliorer la méthode proposée plusieurs voies de recherches sont envisageables.

Parmi celles-ci, des travaux sur la définition de la fonction de coût cible doivent être entrepris. Il pourrait ainsi être judicieux de veiller à mieux satisfaire la cible acoustique en certains points (noyaux vocaliques par exemple) quitte à relâcher les contraintes ailleurs (zones non voisées). Ceci permettrait d'introduire un contrôle acoustique accru là où des cibles vraiment pertinentes et importantes du point de vue de la perception peuvent être définies. Une solution consisterait à pondérer la fonction de coût selon un critère de stabilité par exemple. Une autre façon de procéder reviendrait à rechercher des points cible et de chercher les points qui satisfont le mieux à cette cible. De plus, une telle façon de procéder pourrait s'avérer plus robuste, car basée sur l'exploitation d'une information cible à la fois plus fiable et mieux localisée (c'est-à-dire dont l'impact sur le plan de la perception risque d'être plus important).

Une deuxième perspective de recherche consiste à utiliser une modélisation acoustique plus évoluée que les HMM, de manière à mieux tenir compte des trajectoires temporelles des paramètres acoustiques. En effet, un des principaux inconvénients des HMM est qu'ils supposent, pour un état donné, l'indépendance des observations. De ce fait, ils ne permettent pas réellement de modéliser des trajectoires acoustiques. Il serait donc intéressant de mesurer les performances, en terme de modélisation mais aussi de classification, de méthodes alternatives, telles que les approches à base de modèles de trajectoires (STM pour Stochasric Trajectory Model) [Pel98] ou encore de réseaux dynamiques bayésiens (DBN pour Dynamic Bayesian Network) [Zwe98] [BZ01] [Mur02] [PB05]. Ces techniques de modélisation acoustique ont été appliquées en reconnaissance de la parole, mais n'ont connu jusqu'à présent qu'un succès mitigé. Ceci étant, il faut bien avoir à l'esprit que dans un contexte de reconnaissance de la parole des simplifications de ces modèles ont été faites pour limiter la complexité. En synthèse de la parole, ces contraintes doivent pouvoir être levées car l'objectif est de caractériser des trajectoires acoustiques (éventuellement hors ligne) et non d'effectuer un décodage acoustico-phonétique avec de fortes contraintes en terme de complexité algorithmique. Une piste intéressante serait donc de mesurer l'apport de tels modèles dans un contexte sensiblement moins contraint.

Bibliographie

[Bau72] L.E. Baum. " An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes". *Inequalities*, no. 3, pp. 1–8, 1972.

[BC95] A.W. Black and N. Campbell. "Optimising Selection of Units from Databases for Concatenative Synthesis". *Eurospeech*, September 1995.

[BJ98] A. Breen and P. Jackson. "Non-Uniform Unit Selection and the Similarity Metric Within BT's LAUREATE tts System". *3rd ESCA Int. Workshop*, November 1998.

[Blo03] C. Blouin. "Sélection des unités pour la synthèse vocale par concaténation". PhD thesis, Université Paris-Sud, Décembre 2003.

[BP66] L.E. Baum and T. Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". *Annals of Mathematical Statistics*, pp. 1554–1563, 1966.

[BPQ+99] M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri. "Choose the Best to Modify the Least : A New Generation Concatenative Synthesis System". *Eurospeech*, September 1999.

[BSG+91] L.R. Bahl, P.V. De Souza, P.S. Gopalakrichnan, D. Nahamoo, and M.A. Picheny. "Decision Tree For Phonological Rules in Continuous Speech". *ICASSP*, pp. 185–188, 1991.

[BT97] A.W. Black and P. Taylor. "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis". *Eurospeech*, pp. 601–604, September 1997.

[BZ01] J. Bilmes and G. Zweig. "Discriminatively Structured Dynamic Graphical Models for Speech Recognition "". *Final Report : JHU 2001 Summer Workshop*, 2001.

[CBSB00] A. Conkie, M. C. Beutngel, A. K. Syrdal, and P. E. Brown. "Preselection of Candidate Units in a Unit Selection-Based Test-to-Speech Synthesis System". *ICSLP*, October 2000.

[CI91] N. Campbell and S.D. Isard. "segment durations in a syllable frame". *Special issue on Speech Synthesis*, pp. 37–47, 1991.

[CI96] A. Conkie and S. Israd. "Optimal Coupling of Diphones". *Progress in Speech Synthesis Springer Verlag*, 1996.

[CM98] A. Cronk and M. Macon. "Optimized Stopping Criteria for Tree-Based Unit Selection in Concatenative Synthesis". *ICSLP*, December 1998.

[Com81] P. Combescure. "20 Listes de Dix Phrases Phonétiquement Èquilibrées". *Revue d'Acoustique*, vol. 56, pp. 34–38, 1981.

[Con99] A. Conkie. "Robust Unit Selection for Speech Synthesis". *Joint Meeting of ASA, EAA and DAGA*, March 1999.

[CYC01] M. Chu, H. Yang, and E. Chang. "Selecting non-Uniform Units From a Very Large Corpus for Concatenative Speech Synthesizer". *ICASSP*, 2001.

[DE98] R.E. Donovan and E.M. Eide. "The IBM Trainable Speech Synthesis System". *ICSLP*, 1998.

[DM80] S.B. Davis and P. Mermelstein. "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continously Spoken Sentences". *IEEE Taransactions on Acoustic Speech and Signal Processing*, vol. 28, no 4, pp. 357–366, 1980.

[Don96] R.E. Donovan. *"Trainable Speech synthesis"*. PhD thesis, Cambridge University Engineering Department, Cambridge, England, 1996.

[Don00] R.E Donovan. "Segment Pre-selection in Dicision-Tree Based Speech Synthesis Systems". *ICASSP*, 2000.

[Don01] R.E. Donovan. "A New Distance Measure For Costing Spectral Discontinuities in Concatenative Speech Synthesisers". *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, September 2001.

[Eme77] F. Emerard. *"Synthèse par diphones et traitement de la prosodie"*. PhD thesis, Université de Grenoble III, 1977.

[FF88] G. J. Freij and F. Fallside. "Lexical Stress Recognition Using Hidden Markov Models". *ICASSP*, pp. 135–138, 1988.

[FH82] H. Fujisaki and K. Hirose. "Analysis and Synthesis of Voise Fundamental Frequency Contours of Spoken Sentencesn". *ICASSP*, pp. 950–953, May 1982.

[FKI92] T. Fukada, T. Kobayashi, and S. Imai. "An Adaptive Algorithm for mel-Cepstral Analysis of Speech". *ICASSP*, vol. 1, pp. 137–140, 1992.

[For73] G. Forney. "The Viterbi Algorithm". *Proceedings of the IEEE*, no. 61, pp. 268–278, mars 1973.

[Fra02] H. François. *"Synthèse de la Parole par Concaténations d'Unités Acoustiques : Construction et Exploitation d'une Base de Parole Continue"*. PhD thesis, Université de Rennes I, 2002.

[Fur86] S. Furui. "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum "". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[Fur98] S. Furui. "Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems"". *ICASSP*, pp. 293–296, 1998.

[Gab94] B. Gabioud. "Articulatory Models in Speech Synthesis". *Fundamentals of Speech Synthesis and Speech : Basic Concepts, State of the Art, and Future Challenges Recognition*, pp. 215–230, 1994.

[GH87] Y. Gong and J. Haton. "Time Domain Harmonic Matching Pitch Estimation Using Time-dependent Speech Modeling". *IEEE Trans. Acoust., Speech, Signal processing*, vol. 35, no. 10, pp. 1386–1400, 1987.

[HAA+96] X. Huang, A. Acero, J. Adcock, H. Hon, J. Goldsmith, J. Liu, and M. Plumpe. "Whistler, a trainable textto-speech system". *ICSLP*, vol 4, pp. 2387–2390, October 1996.

[HAH+97] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe. "Recent Improvements on Microsoft's Trainable Text-to-Speech System : Whistler ". *IEEE ICASSP*, pp. 959–962, 1997.

[HPAA99] X. Huang, M. Plumpe, A. Acero, and J. Adcock. "Method and System for Runtime Acoustic Unit Selection for Speech Synthesis ". *United States Patent*, no. 913193, 1999.

[Har53] C.M. Harris. "A study of The Building Blocks of Speech". *Journal of the Acoustical Society of America*, vol. 25, no 5, pp. 962–969, 1953.

[HB96] A. Hunt and A. Black. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database". *ICASSP*, May 1996.

[HC98] M. Holzapfel and N. Campbell. "A Nonlinear Unit Selection Strategy for Concatenative Speech Synthesis Based on Syllable Level Features". *ICSLP*, December 1998.

[Her88] D. Hermes. "Measurement of Pitch by Subharmonic Summation". *Journal of the Acoustical Society of America*, pp. 257–264, 1988.

[Hes83] W. Hess. "Pitch Determination of Speech Signal". *Springer-Verlag*, 1983.

[Hir89] T. Hirokawa. "Speech Synthesis Using a Waveform Dictionary". *EUROSPEECH*, pp. 140–143, September 1989.

[HMC89] C. Hamon, E. Moulines, and F. Charpentier. "A Diphone Synthesis System Based on Time-Domain Prosodic Modification of Speech ". *ICASSP*, pp. 238–341, 1989.

[IKS92] N. Iwahashi, N. Kaiki, and Y. Sagisaka. "Concatenative Speech Synthesis by Minimum Distortion Criteria". *ICASSP*, 2, March 1992.

[KK04] R. Kumar and S. P. Kishore. "Automatic Pruning of Unit Selection Speech Databases for Synthesis without loss of Naturalness". *ICSLP*, 2004.

[KL51] S. Kullback and R. A. Leibler. "Information and Sufficiency ". *Annals of Mathematical Statistic*, vol. 22, pp. 76–86, 1951.

[Kla79] D.H Klatt. "Synthesis by Rule of Segmental Duration in English Sentences". *Frontiers of Speech Communication Research*, pp. 287–299, 1979.

[KV98] E. Klabbers and R. Veldhuis. "the Reduction of Concaténation Artefacts in Diphone Synthesis ". *ICSLP*, pp. 1983–1986, 1998.

[LC98] Y. Laprie and V. Colotte. "Automatic Pitch Marking for Speech Transformations via TD-PSOLA ". *EUSIPCO*, 1998.

[LHSW04] Z. H. Ling, Y. Hu, Z. W. Shuang, and R. H. Wang. "Compression of Speech Data by Feature Separation and Patten Clustering Using STRAIGHT". *ICSLP*, 2004.

[LLO01] M. Lee, D.P. Lopresti, and J.P. Olive. "A Text-to-Speech Platform for Variable Length Optimal Unit Searching Using Perceptual Cost Functions". *4th ISCA Tutorial and Research Workshop on Speech Synthesis.*, September 2001.

[LSM93] J. Laroche, Y. Stylianou, and E. Moulines. "Speech Modification Based on a Harmonic + Noise Model ". *ICASSP*, 1993.

[Mae79] S. Maeda. "Un Modèle Articulatoire de la Langue avec des Composantes Lineaire ". *10 èmes Journées d'Etude sur la Parole*, pp. 152–162, 1979.

[MC90] E. Mouline and F. Charpentier. "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech synthesis using diphones.". *Speech Communication*, vol. 9, pp. 453–467, 1990.

[Meu96] P.Y. LE Meur. *"Synthèse de Parole par Unités de Taille Variable"*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[MO86] I. Mikuni and K. Ohta. "Phoneme Based Text-to-Speech Synthesis System". *ICASSP*, pp. 2435–2438, April 1986.

[MQ90] R.J. McAulay and T.F. Quatieri. "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speechmodel". *IEEE Trans. Acoust., Speech, Signal processing*, vol. 1, pp. 249–252, 1990.

[MTKI96] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. "Speech Synthesis Using HMMs With Dynamic Features". *ICASSP*, pp. 389–392, 1996.

[Mur02] K. P. Murphy. *"Dynamic Bayesian Networks : Representation, Inference and Learning"*. PhD thesis, Computer science, University of California, Berkeley, 2002.

[NH88] S.Y. Nakajima and H. Hamada. "Automatic Generation of Synthesis Units Based on Context Oriented Clustering". *ICASSP*, vol. 115, pp. 659–662, April 1988.

[NMS90] T. Nomura, H. Mizuno, and H. Sato. "Speech Synthesis by Optimum Concatenation of Phoneme Segments". *1st ESCA-IEEE Tutorial and Research Workshop on Speech Synthesis*, pp. 39–42, 1990.

[Ode95] J.J. Odell. "The Use of Context in Large Vocabulary Speech Recognition". PhD thesis, Queen' College, March 1995.

[PA01] R. Prudon and C. Alessandro. "A Selection/Concatenation Test-to-Speech System : Databases Developement, System Design, Comparative Evaluation ". *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, September 2001.

[PB05] F. Pernkopf and J. Bilmes. "Discriminative versus Generative Parameter and Structure Learning of Bayesian Network Classifiers "". *International Conference on Machine Learning, Bonn, Germany*, 2005.

[Pel98] B. Pellom. *"Enhancement, Segmentation, and Synthesis of Speech with Application to Robust Speaker Recognition"*. PhD thesis, Electrical engineering, Duke University, 1998.

[Pie80] J. Pierrehumber. "The Phonology and Phonetics of English Intonation". PhD thesis, MIT, Boston, 1980.

[PSK05] Y. Pantazis, Y. Stylianou, and E. Klabbers. "Discontinuity Detection in Concatenated Speech Synthesis Based on Nonlinear Speech Analysis ". *INTERSPEECH*, 2005.

[PWOB90] P.J. Price, C.W. Wightman, M. Ostendorf, and J. Bear. "The use of Relative Duration in Syntactic Disambigution ". *Internatinal Conference on Spoken Language*, vol. 1, pp. 13–16, 1990.

[PZC02] H. Peng, Y. Zhao, and M. Chu. "Perceptually Optimizing the Cost Function for Unit Selection in TTS System With one Single Run of MOS Evaluation". *ICSLP*, pp. 2613–2616, September 2002.

[RAFT02] P. Rutten, M. Aylett, J. Fackrell, and P. Taylor. "A Statistically Motiveted Batabase Pruning Technique for Unit Selection Synthesis ". *ICSLP*, 2002.

[RJ89] L.R Rabiner and B.H. Juang. "Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[RJ93] L.R. Rabiner and B.H. Juang. " Fundamentals of Speech Recognition". édition Prentice Hall PTR, 1993.

[RO94] K. Ross and M. Ostendorf. "A Dynamical System Model for Generating F_0 for Synthesis". *ESCA/IEEE Workshop on Speech Synthesis*, pp. 131–134, 1994.

[Sag88] Y. Sagisaka. "Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units". *ICASSP*, April 1988.

[SK94] T. Styger and E. Keller. "Formant synthesis". *Fundamentals of Speech Synthesis and Speech : Basic Concepts, State of the Art, and Future Challenges*, pp. 109–128, 1994.

[SKIM92] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. "ATR v-Talk Speech Synthesis System". *International Conference on Spoken Language Systems*, pp. 483–486, October 1992.

[SS01] Y. Stylianou and A. K. Syrdal. "Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis ". *ICASSP*, 2001.

[SSF87] P.L. Salza, S. Sandri, and E. Foti. "Evaluation of Experimental Diphones for Text-to-Speech Synthesis of Italian". *Eurospeech*, pp. 63–66, September 1987.

[Sty96] Y. Stylianou. "Harmonic Plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modifications". PhD thesis, Telecom Paris, Janvier 1996.

[SYK01] K. Sanghun, L. Youngjik, and H. Keikichi. "Pruning of Redundant Synthesis Instances Based on Weighted Vector Quantization". *Eurospeech*, 2001.

[Tay93] P. Taylor. "ATR Automatic Recognition of Intonation From F0 Contours Using the Rise/Fall/Connection Model". *Eurospeech*, vol. 5, pp. 789–792, 1993.

[TFBH94] J. Taboada, S. Feijoo, R. Balsa, and C. Hernandez. "Explicit Estimation of Speech Boundaries". *IEEE Proc-Sci. Meas. Technol.*, vol. 141, no. 3, pp. 123–126, 1994.

[TKTS02] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. "Unit Selection for Japanese Speech Synthesis Based on Both Phoneme Unit and Diphone Unit". *ICASSP*, vol. 1, pp. 465–468, May 2002.

[TMMK99] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling". *ICASSP*, vol. 1, pp. 229–232, 1999.

[Tou98] S. De Tournemire. "Identification et Génération Automatique de Contours Prosodiques pour la Synthèse Vocale à Partir du Texte en Français". PhD thesis, Ecole Nationale Supérieure des Télécommunication, Paris, Avril 1998.

[TZB02] K. Tokuda, H. Zen, and A.W. Blak. "An HMM-Based Speech Synthesis System Applied To English". *Workshop on TTS*, vol 3, pp. 2263–2266, September 2002.

[Vit67] A.J. Viterbi. "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm". *IEEE Transactions on Information Theory*, vol. IT-13, pp. 260–269, 1967.

[vSB97] J. P. H. van Santen and A. L. Buchsbaum. "Methods for optimal text selection". *EUROSPEECH*, pp. 553–556, 1997.

[WCI93] W.J. Wangt, W.N. Campbell, and N. Iwahashi. "Tree-Based Unit Selection for English Speech Synthesis". *IEEE ICASSP*, vol. II, pp. 191–194, April 1993.

[WTTS98] *Workshop on Text-to-Speech Synthesis.*, 1998.

[YG98] G.R.W Yi and J. Glass. "Natural-Sounding Speech Synthesis Using Variable-Length Units". *ICSLP*, pp. 2617–2620, November-December 1998.

[YG02] G.R.W Yi and J. Glass. "Information-Theoretic Criteria for Unit Selection Synthesis". *ICSLP*, pp. 2617–2620, September 2002.

[YWZ+04] Z. L. Yu, K. Z. Wang, Y. Q. Zu, D. J. Yue, and G. L. Chen. "Data Pruning Approach to Unit Selection for Inventory Generation of Concatenative Embeddable Chinese TTS Systems". *ICSLP*, 2004.

[ZCPC03] Y. Zhao, M. Chu, H. Peng, and E. Chang. "Custom-Tailoring TTS Voice Font-Keeping the Naturalness When Reducing Database Size". *Eurospeech*, 2003.

[Zwe98] G. Zweig. *"DBN Based Speech Recognition"*. PhD thesis, U.C. Berkeley, 1998.

Annexe A

Modèles HMMs

Les HMMs ont été utilisés de manière intensive en reconnaissance automatique de la parole [RJ89]. Dans ce domaine, les signaux sont codés comme des variations temporelles de spectres de courte durée. L'application des HMMs s'étend maintenant à des domaines tels que la reconnaissance des formes, le traitement du signal et la synthèse de parole. Un HMM est un double processus stochastique dont un processus sous-jacent est non observable mais peut être estimé à partir d'un ensemble de processus qui produisent une séquence d'observations. Les HMMs peuvent être utilisés pour le traitement de problèmes dans lesquels l'information est incertaine et incomplète. Leur utilisation nécessite deux étapes : une étape d'apprentissage au cours de laquelle le processus stochastique est estimé à partir d'observations extensives et une étape de mise en oeuvre où le modèle peut être utilisé en temps réel pour obtenir les séquences de probabilités maximales. Les modèles de Markov cachés sont robustes et fiables du fait de l'existence de nombreux algorithmes d'apprentissage efficaces et robustes.

A.1 Présentation

Un modèle de Markov caché ou HMM est donc un modèle stochastique particulier, représentant une séquence par deux suites de variables aléatoires, l'une étant cachée et l'autre observable :

– La suite cachée correspond à la suite des états $q_1...q_T$, notée $Q(1:T)$, où les q_i prennent leur valeur parmi l'ensemble des n états du modèle $\{s_1...s_n\}$;

– La suite observable correspond à la séquence d'observations $o_1...o_T$, notée $O(1 : T)$, où o_i sont aussi fonctions du temps et se réalisent parmi un ensemble de M symboles observables $\{v_1...v_m\}$.

En pratique, les HMM construits sont ceux pour lesquels les suites observables sont les séquences que l'on cherche à modéliser. Les séquences observées peuvent alors être définies comme des phrases sur l'alphabet$\{v_1...v_m\}$. Un modèle de Markov caché est principalement défini par deux matrices et un vecteur :

– une matrice A de *probabilités de transition* entre les états de la chaîne. a_{ij} représente la probabilité que le modèle évolue de l'état i vers l'état j :

$$a_{ij} = A(i,j) = P(q_{t+1} = s_j | q_t = s_i), \qquad \forall i,j \in [1...N] \forall t \in [1...T] \tag{A.1}$$

avec :

$$\forall i,j : a_{ij} \geq 0 \qquad \text{et} \qquad \forall i : \sum_{j=1}^{n} a_{ij} = 1 \tag{A.2}$$

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

$$b_j(k) = P(o_t = v_k | q_t = s_j), \qquad 1 \leq j \leq n, 1 \leq k \leq M \tag{A.3}$$

– un vecteur $\pi = \{\pi_i\}_{i=1..n}$ de *densités de probabilité initiale.*π_i représente la probabilité que l'état de départ du modèle soit l'état i, soit :

$$\pi_i = P(q_1 = s_i), \qquad 1 \leq i \leq n \tag{A.4}$$

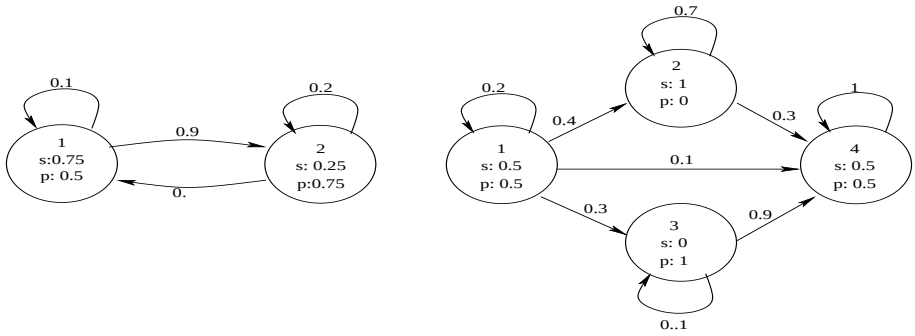
avec :

$$\forall i : \pi_i \geq 0 \qquad \text{et} \qquad \sum_{i=1}^{n} \pi_i = 1 \tag{A.5}$$

Il existe deux types principaux de modèles de Markov cachés, le modèle *ergodique* et le modèle *gauche-droite* :

– Le modèle ergodique est un modèle sans contrainte où toutes les transitions d'un état vers un autre sont possibles.

– Le modèle gauche-droite contient au contraire des contraintes : il y a interdiction de certaines transitions par la mise à zéro des valeurs a_{ij} correspondantes figure A.1.



(a) HMM ergodique à deux états.

(b) HMM gauche−droit à quatre états.

Figure A.1 – Deux exemples d'HMM

La parole étant un phénomène temporel, l'utilisation de HMMs pose comme postulat que la parole est une suite d'évènements stationnaires. La topologie principalement employée dans la littérature est un modèle gauche-droit d'ordre 1 dit de Bakis figure A.2. Cette topologie permet la modélisation des variations temporelles au sein du signal de parole. Le lecteur trouvera de plus amples informations sur les HMM dans [RJ89] et [RJ93].

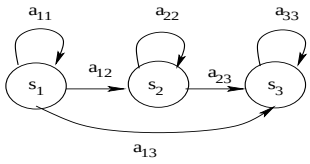


Figure A.2 – Modèle HMM dit gauche-droit d'ordre 1 à 3 états

A.2 Apprentissage des modèles HMMs

L'un des principaux problèmes de l'utilisation des HMMs réside dans la phase d'apprentissage, qui conduit à l'évaluation de tous les paramètres du modèle. Il s'agit avec un corpus d'apprentissage, contenant un étiquetage par sous-unités acoustiques du signal temporel, de maximiser la vraisemblance que le modèle HMM ait produit la suite d'observations. Il existe plusieurs algorithmes pour faire cela : l'algorithme dit de Baum-Welch [BP66], le forward-backward [Bau72] ou même simplement à l'aide de l'algorithme de Viterbi [For73]. Ces techniques d'apprentissage des modèles acoustiques

n'étant pas directement reliées à notre thèse, nous ne détaillerons pas ces algorithmes dans ce manuscrit. Pour de plus amples informations sur ces algorithmes, le lecteur pourra se référer à [RJ89] et [RJ93].

A.3 Arbres de décision

La modélisation HMM est réalisée, généralement, dans le but de rendre plus facile la classification des unités modélisées. Les arbres de décision sont les techniques de classification les plus répandues [Ode95],[Don96].

Le principe d'une telle technique est d'organiser l'ensemble des données comme un arbre : une feuille de cet arbre désigne une des C classes (à chaque classe peuvent correspondre plusieurs feuilles) et à chaque nœud interne est associé un test portant sur un ou plusieurs éléments de l'espace de représentation. La réponse à ce test désignera le fils du nœud vers lequel on doit aller. La classification s'effectue donc en partant de la racine pour poursuivre récursivement le processus jusqu'à ce que l'on rencontre une feuille.

Dans le cadre d'une modélisation par HMM, les unités à classifier sont les états de ces modèles (les sénones). Les classes peuvent être de différents types mais sont très souvent de nature acoustique. Nous citerons, à titre d'exemple, les plosives, les fricatives, les liquides, les voisées et les non voisées mais il en existe une multitude d'autres. La classification est accomplie en se basant sur les questions sur le contexte gauche et droit des phonèmes. Les deux critères d'arrêt sont le nombre minimum de trames de la parole qui doivent être assignées à chaque nœud et l'augmentation minimum de probabilité qui doit être réalisée pour que le nœud soit dédoublé. Les nœuds terminaux finaux forment les états groupés pour chaque arbre.