UNIVERSITE SAAD DAHLAB DE BLIDA

Faculté des Sciences de l'Ingénieur

Département d'Electronique

Thèse de Doctorat d'Etat

Spécialité : Traitement de la parole

SYNTHESE DE LA PAROLE DE L'ARABE STANDARD

Par

Fatima CHOUIREB

devant le jury composé de :

A. Guessoum	Professeur, U.S.D, Blida	Président
M. Guerti	Maître de conférences, ENP, Alger	Rapporteur
H. Salhi	Maître de conférences, U.S.D, Blida	Examinateur
M. Halimi	Maître de recherche, C.S.C, Cheraga	Examinateur
H. Sayoud	Maître de conférences, U.S.T.H.B, Alger	Examinateur

Blida, décembre 2007

ملخص
إن مجال معالجة الكلام آليا يلقى في هذه الأيام اهتماما متزايدا من طرف الباحثين لأنه يمّكن الإنسان من التواصل مع الآلة لتقوم بكثير من الأعمال بدلا منه. و يتمثل هذا التواصل في نقطتين رئيسيتين: تتمثل الأولى في تمكين الآلة من النطق بطريقة صحيحة والثانية في التعرف الآلي على الكلام. و يعتبر التحويل الآلي لأي نص مكتوب إلى كلام مفهوم و طبيعي منطوق بواسطة نظام آلي مهمة معقدة تتطلب المرور بعدة مراحل و تحليلات لغوية و صوتية. و كان هذا التحويل يعتمد أساسا على إحدى الطريقتين التاليتين: الأولى تتمثل في تركيب الكلام باستعمال عدة قواعد يقوم بجمعها و استنباطها لغويون و نحاة ضالعون في اللغة. و الطريقة الثانية تعتمد على تركيب الكلام عن طريق وحدات صوتية مسجلة مسبقا لتكوين الجمل المطلوبة، و كلتا الطريقتين لهما مساوئ من بينها : صعوبة تكوين القواعد بالنسبة للطريقة الأولى و الذاكرة الكبيرة المطلوبة لتخزين الوحدات الصوتية اللازمة في تسلسل متنوع بالإضافة إلى الصعوبة في تركيب هذه الوحدات بدون وقوع خلل سمعي عند المرور من صوت إلى آخر. في هذا العمل ارتأينا استعمال طريقة أخرى تعتمد على تدريب الآلة على النطق باستعمال الشبكة العصبية. و هذا التدريب يتطلب تحليلا لفظيا و نبريا لمتن صوتي مسجّل ومصنّف. ونظرا لقلة الأعمال في هذا الميدان بالنسبة للغة العربية، نقدم أيضا في هذا العمل المراحل التي اتبعناها لإنجاز قاعدة معلومات صوتية نبرية للغة العربية كتلك المتوفرة للغات الأجنبية كقاعدة المعلومات TIMIT للإنجليزية و BDSONS للفرنسية بالإضافة أيضا إلي مولد آلي للنبرة ليكون الكلام المنطوق طبيعيا و ذو جودة عالية.

الكلمات المفاتيح تركيب نص-إلى- كلام، التقطيع الى فونيمات و صوتيات، الشبكات العصبية، معالجة اللغة الطبيعية، قاعدة معلومات نبرية، مركب معلومات نبرية.

Résumé La synthèse de la parole à partir du texte TTS (Text-To-Speech) est une tâche complexe qui vise à convertir un texte écrit quelconque en parole intelligible et naturelle afin de communiquer l'information d'une machine à l'homme. Les deux méthodes classiques qui permettent de faire cette conversion sont la synthèse par règles et la synthèse par concaténation d'unités pré-stockées. Ces méthodes montrent certains inconvénients tels que : l'effort considérable pour l'établissement des règles de transitions entre les différentes représentations phonétiques dans la synthèse par règles et la difficulté dans la sélection des meilleurs segments à partir d'une large base de données ainsi que l'espace mémoire nécessaire pour le stockage de ces segments dans la synthèse par concaténation. Par conséquent, il y a toujours un besoin de rechercher des méthodes de synthèse améliorées pour accroître la qualité et le naturel de la parole générée et diminuer l'espace mémoire requis pour le bon fonctionnement des systèmes TTS. Cette thèse décrit les différentes étapes de développement d'un système de synthèse TTS pour l'Arabe Standard à base de Réseaux de Neurones (RN) et d'un modèle de Prédiction Linéaire à Excitation Résiduelle RELP (Residual Excited Linear Prediction). Ces réseaux de neurones nécessitent de larges bases de sons préalablement étiquetées prosodiquement. Vu le manque d'une telle base de données pour la langue Arabe, nous avons réalisé une comme celles disponibles pour les autres langues (la base de données TIMIT pour l'Anglais et BDSONS pour le Français, etc.). Ainsi, nous avons discuté les différentes étapes entreprises pour l'élaboration de cette base de données. Un synthétiseur des informations prosodiques à base de réseaux de neurones est également décrit dans ce travail.

Mots clés Synthèse TTS ; Segmentation acoustique et phonétique ; Réseaux de neurones ; Traitement du langage ; naturel ; Excitation résiduelle ; Base de données prosodique ; synthétiseur des informations prosodiques.

Abstract Text–to–speech (TTS) synthesis is a complex task, whereby an unknown input text is converted into the spoken word, via a series of linguistic and speech analyses. This conversion has traditionally been performed either by concatenating short samples of speech or by using rule-based systems to convert a phonetic representation of speech into an acoustic representation, which is then converted into speech. This thesis describes a text-to-Speech synthesis system for modern standard Arabic based on artificial neural networks (ANNs) and Residual Excited Linear Prediction (RELP) coder. The networks offer a storage-efficient means of synthesis without the need for explicit rule enumeration. These neural networks require large prosodically labelled continuous speech databases in their training stage. As such databases are not available for the Arabic language; we have developed one for this purpose. Thus, we discuss various stages undertaken for this development process. In addition to interpolation capabilities of neural networks, a linear interpolation of the coder parameters is performed to create smooth transitions at segment boundaries. A residual-excited all-pole vocal tract model and a prosodic-information synthesizer based on neural networks are also described in this work.

Keywords Text-To-Speech synthesis, phonetic and acoustic segmentation, neural networks, natural language processing, residual excitation, prosodic database, prosodic-information synthesizer.

REMERCIEMENTS

Au terme de ce travail, je tiens tout d'abord à remercier vivement mon encadreur Madame Mhania Guerti qui m'a offert la possibilité de réaliser une thèse de doctorat sous sa direction. Je la remercie également pour ses discussions fructueuses, ses encouragements, ses conseils judicieux et ses suggestions. Qu'elle trouve ici mes meilleurs sentiments de gratitude.

Je remercie vivement Monsieur Abderazak Guessoum, Professeur à l'université de Blida, qui m'a fait l'honneur d'accepter de présider le jury et de juger ce travail.

Je tiens à exprimer ma reconnaissance à Monsieur Hassan Salhi Maître de conférences à l'université de Blida, à Monsieur Mohamed Halimi, Maître de recherche au Centre de Recherche Scientifique et Technique en Soudage et Contrôle CSC, à Monsieur Halim Sayoud, Maître de Conférences à l'Université des Sciences et Technologies Houari Boumediene (USTHB) d'avoir bien voulu juger ce travail et participer au jury.

Je tiens également à exprimer mes remerciements à mes collègues et amies à l'université de Laghouat et l'université de Blida. Je pense en particulier à B. Yousfi, A. Choucha, N. Benblidia, F. Z. Reguieg, F. Doudou, A. Chentir, K. Kouzi, et S. Bellakhal. Mes remerciements vont également à tout le personnel de l'université de Laghouat, enseignants et administrateurs, en particulier ceux du Département de Génie Electrique.

Je ne saurais terminer ces remerciements sans mentionner les membres de ma petite et ma grande famille et spécialement ma mère et mon mari, qui, sur le plan humain, m'ont soutenu par leurs encouragements tout au long de la réalisation de ce travail. Qu'ils trouvent ici toute ma gratitude et mon amour pour eux.

Que tous ceux qui ont contribué, de près ou de loin, à la réalisation de cette thèse trouvent ici mes meilleures gratitudes.

TABLE DES MATIERES

LISTES DES FIGURES

LISTES DES TABLEAUX

INTRODUCTION GENERALE

La synthèse de la parole à partir du texte TTS (Text-To-Speech) désigne l'ensemble des traitements permettant à une machine de convertir un texte écrit quelconque en message oral. Les systèmes capables de réaliser cette tâche sont actuellement très demandés dans diverses applications, tels que les machines de lecture pour les aveugles, les machines parlantes pour les gens qui ont perdu l'utilisation de leur voix, et les dispositifs pour l'accès à l'information électroniquement stockée par téléphone (lecture à distance d'email, les services automatiques de réservation de ligne aérienne, etc.). Des applications existent également dans l'enseignement des langues, où les systèmes TTS pourraient démontrer la prononciation correcte des mots et des phrases arbitraires. D'autres possibilités incluent la sortie parole des systèmes de translation automatique et des machines intelligentes.

Le nombre d'applications possibles de la technologie TTS a considérablement augmenté pendant les années récentes, et ceci en grande partie, en raison de l'augmentation rapide de l'utilisation des ordinateurs dans la société. Dès lors, les travaux des chercheurs se sont dirigés, de plus en plus, vers l'amélioration de l'intelligibilité et du naturel des systèmes TTS, afin d'atteindre une meilleure qualité, sur le timbre de la voix synthétique ainsi que sur son élocution et son intonation [1-3].

Le développement d'un système TTS à vocabulaire illimité est une tâche énorme avec des difficultés surgissant à chaque étape du processus. Les deux méthodes classiques qui permettent de faire cette conversion sont la synthèse par règles [4] et celle par concaténation d'unités acoustiques issues d'une parole naturelle pré-enregistrée [5, 6]. Cette dernière catégorie produit une parole synthétique intelligible. Pour obtenir une parole naturelle, un traitement supra-segmental doit lui être appliqué. Certains de ces systèmes utilisent des inventaires d'unités de taille fixe, tels que les diphones ou les triphones avec un seul exemplaire de chaque type. D'autres systèmes utilisent des unités de taille variable (non uniformes) et multireprésentées extraites d'un grand corpus renfermant plusieurs

représentations de chaque unité. La grande variabilité dans de tels segments acoustiques naturels permet une haute qualité de la parole, une bonne modélisation du naturel, et des différences dans les styles parlés [7]. Cependant, la concaténation directe des segments extraits d'un grand corpus est coûteuse en termes de collection des données, des exigences des méthodes de sélection des unités, de l'organisation et de l'espace mémoire nécessaire pour le stockage de la base de données des segments de la parole. Les systèmes de synthèse par règles sont également utilisés pour convertir des représentations phonétiques en signal vocal. Ils stockent des valeurs cibles pour chaque représentation phonétique possible. Ces valeurs sont modifiées selon un ensemble de règles de transition entre les différentes représentations phonétiques. La méthode de synthèse par règles a comme principal inconvénient le fait de recourir à un nombre important de règles de transition pour modéliser au mieux les caractéristiques de la parole humaine. La pertinence et le nombre des valeurs cibles est également un élément important, car une représentation inadéquate ou trop simpliste du signal de la parole peut dégrader la qualité globale de la synthèse [8, 9, 10].

De nos jours, l'utilisation des méthodes basées sur l'analyse de corpus de la parole est devenue la méthodologie primaire dans les domaines de la synthèse et la Reconnaissance Automatique de la Parole (RAP). Ces techniques qui se basent sur l'apprentissage automatique de la machine doivent leur efficacité et leur performance aux développements de plus en plus d'ordinateurs performants et des modèles de calculs puissants, tels que les Réseaux de Neurones (RN) ou les Modèles de Markov Cachés (HMM).

Les réseaux de neurones ont le potentiel de générer une parole synthétique plus naturelle que les autres technologies de synthèse, car ils sont entraînés à partir de signaux de parole naturelle. Ils exigent moins d'espace mémoire et moins d'effort manuel que les techniques classiques. Ils ont aussi la capacité de produire des variations temporelles à l'intérieur du segment phonétique et des différentes sorties quand les contextes d'entrée changent. En outre, l'architecture distribuée des réseaux leur permet de généraliser les connaissances extraites de l'ensemble d'apprentissage à de nouvelles situations [11-16]. Ces propriétés permettent à un synthétiseur neuronal de la parole de modéliser les effets de coarticulation dans une large gamme de contextes en utilisant une quantité limitée de données d'apprentissage.

Dans un système TTS, l'information prosodique joue un rôle important dans l'intelligibilité et le naturel de la parole générée. Vu la puissance croissante des ordinateurs

modernes, il y a eu un intérêt grandissant pour la génération de la prosodie par réseaux de neurones. Au moment où des générateurs de prosodie de haute performance ont été développés pour les autres langues [17-22], peu de travaux ont été faits pour générer la prosodie de l'Arabe Standard.

Notre travail propose l'utilisation des réseaux de neurones multicouches à rétro-propagation du gradient et du codeur LPC à Excitation Résiduelle RELP (Residual Excited Linear Predictor), pour le développement d'un système de synthèse de la parole à partir du texte Arabe Standard (AS). Notre investigation a essentiellement porté sur le réseau de neurone phonétique-acoustique qui convertit les informations linguistiques et la durée segmentale en paramètres LSF nécessaires pour générer le signal vocal. Un synthétiseur des informations prosodiques à base de réseaux de neurones a également été développé pour assurer le naturel de la parole générée. Dans l'ordre de faire l'apprentissage de tous les réseaux de neurones utilisés dans ce travail, une Base de Données (BD) prosodiques de l'Arabe standard a été élaborée. Quelques solutions pratiques sont proposées pour le processus de développement de cette base de données. Enfin, un modèle d'excitation résiduelle est appliqué pour améliorer la qualité de la parole synthétique générée.

Cette thèse est organisée en cinq chapitres :

- le premier commence par des généralités sur la production de la parole et sur le système phonétique de l'Arabe standard. Ensuite, les différentes étapes du processus de synthèse de la parole à partir du texte sont examinées, en se concentrant essentiellement sur la génération du signal vocal. Une comparaison des trois principales approches de synthèse a été faite en dégageant leurs avantages et inconvénients. Ce qui prouve qu'il y a toujours un besoin de nouvelles techniques de synthèse plus améliorées ;
- le deuxième présente une description des paramètres prosodiques et une revue des principaux modèles de génération de la prosodie ;
- le troisième illustre l'implémentation des quatre modèles de parole susceptibles d'apporter la haute qualité recherchée, à savoir les modèles LPC, TD-PSOLA, LP-PSOLA, et le modèle HNM ;
- le quatrième décrit les différentes étapes entreprises pour le développement d'une base de données prosodiques pour l'Arabe standard ;

- le cinquième concerne la description détaillée du système TTS proposé, en mettant l'accent sur : l'architecture et l'apprentissage du réseau de neurones phonétique-acoustique, la technique de lissage spectrale des coefficients LSF aux points de concaténations entre phonèmes adjacents, les réseaux de neurones utilisés pour la synthèse des informations prosodiques, et le modèle d'excitation résiduelle. Les commentaires des résultats expérimentaux, les tests d'écoute subjectifs ainsi que l'analyse statistique sont également présentés.

  Enfin, nous terminons notre travail par des conclusions et perspectives.

CHAPITRE 1

NOTIONS FONDAMENTALES SUR L'ARABE STANDARD ET SYNTHESE DE LA PAROLE A PARTIR DU TEXTE

1.1. Introduction

La synthèse de la parole à partir du texte (TTS : Text-To-Speech) est une tâche complexe qui vise à convertir un texte écrit quelconque en parole intelligible et naturelle afin de communiquer l'information d'une machine à l'homme. Malgré les avancées réalisées ces dernières années dans ce domaine, des progrès restent à faire pour accroître la qualité et le naturel de la parole générée et diminuer l'espace mémoire requis pour le bon fonctionnement des systèmes TTS. La réalisation d'un système de synthèse de la parole à partir du texte Arabe Standard (AS) et l'amélioration de certaines de ses composantes est une motivation derrière ce travail.

Pour comprendre comment les différentes méthodes de synthèse et d'analyse de la parole fonctionnent, nous devons avoir quelques connaissances et quelques notions sur la production de la parole, et sur la phonétique de l'Arabe standard sachant qu'un système TTS dépend fortement de la langue étudiée. La théorie de base de ces matières va être discutée brièvement, ensuite nous décrivons les différentes étapes du processus de synthèse de la parole à partir du texte.

1.2. Phonétique de l'AS et théorie de la production de la parole

Les gens ont besoin de langage pour communiquer entre eux. Ce langage est un ensemble de signaux de la parole composés de petites unités appelées phonèmes. L'Arabe standard est la langue de référence qui est utilisée officiellement dans l'enseignement, la littérature, l'administration et les médias dans l'ensemble du monde Arabe. La première étude sur la phonétique de la langue Arabe a été faite par Sibawayh qui montre le mécanisme et l'articulation de chaque phonème, et clarifie l'anatomie du conduit vocal.

Ces dernières années, l'Arabe connaît un regain d'intérêt, notamment dans le domaine de traitement automatique de la parole. Des études concernant l'Arabe Standard ont été faites par plusieurs chercheurs, nous citons les travaux d'EL-ANI sur la phonétique et la phonologie de l'Arabe [23], et les travaux de M. GUERTI [24], de S. BALOUL [10], et de Y. EL-IMAM [25] sur ses particularités et son traitement automatique. Nous présentons par la suite les mécanismes de production de la parole et certaines caractéristiques phonétiques de l'Arabe standard qui permettront de comprendre nos implémentations futures.

1.2.1. Mécanismes de la production de la parole

Le processus de production de la parole est un mécanisme très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique [26]. Une grande quantité d'organes et de muscles entrent en jeu dans la production des sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose sur l'interaction entre trois entités : les poumons, le larynx, et le conduit vocal (figure 1.1).

Le conduit vocal s'étend des cordes vocales jusqu'aux lèvres dans sa partie buccale et jusqu'aux narines dans sa partie nasale. La forme de ce conduit, déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du palais, détermine le timbre des différents sons de la parole. Le conduit vocal est ainsi considéré comme un filtre pour les différentes sources de production de parole telles que les vibrations des cordes vocales ou les turbulences engendrées par le passage de l'air à travers les constrictions au niveau du conduit vocal. Ce filtre est habituellement composé d'un certain nombre de résonances, appelés *formants* et de temps en temps des anti-résonances ou antiformants.

La parole est produite en excitant les résonances et anti-résonances du filtre du conduit vocal. Le son résultant peut être classé comme voisé (ou sonore) ou bien non voisé (ou sourd) selon que l'excitation vient des vibrations des cordes vocales, à une fréquence fondamentale F_0, ou d'un bruit turbulent créé au niveau d'une constriction quelque part dans le conduit vocal. Dans certains sons les deux types d'excitation sont présents en même temps.

1.2.2. Généralités sur le système phonétique de l'Arabe standard

Le système phonétique de l'Arabe Standard (AS) comprend 34 phonèmes dont 28 consonnes, 3 voyelles brèves et 3 voyelles longues. L'Arabe standard est la langue des

médias et de l'enseignement. Elle a la particularité de posséder des consonnes emphatiques, pharyngales et laryngales en plus des modes d'articulation communs à d'autres langues. Chaque consonne est suivie par une voyelle brève ou longue, mais on peut trouver des consonnes non suivies de voyelles, ce qui est connu en Arabe par le sukuun. Pour la transcription phonétique d'un texte Arabe, nous avons choisi le système de transcription SAMPA (Speech Assessment Methods Phonetic Alphabet). Le tableau 1.1 montre les consonnes Arabes et leurs notations phonétiques SAMPA alors que le tableau 1.2 montre les voyelles ainsi que leurs symboles phonétiques SAMPA. Notons que les consonnes emphatiques en Arabe sont : s., d., t., z. et q.

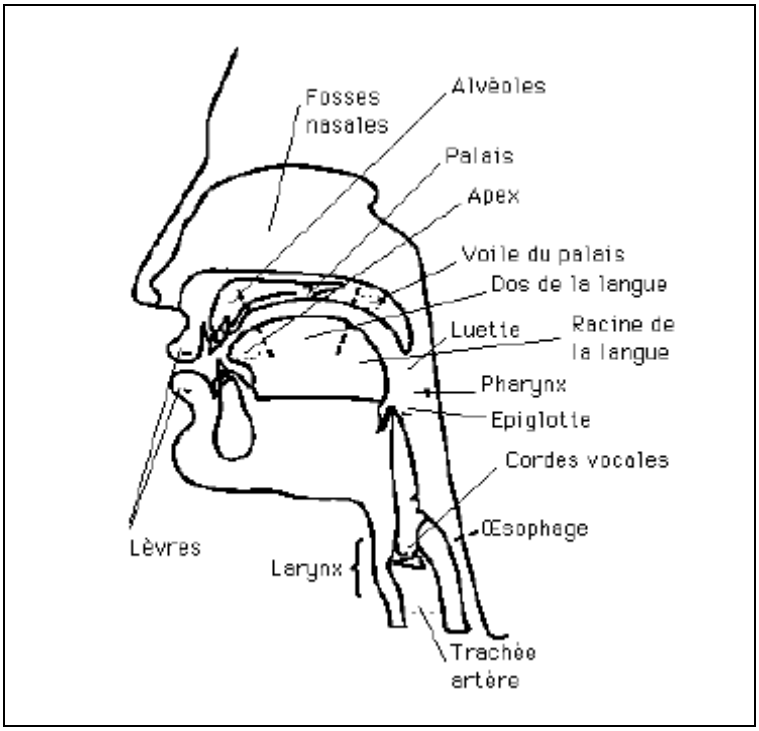


Figure 1.1 : Appareil phonatoire humain

1.3. Synthèse de la parole à partir du texte

L'objectif de la synthèse de la parole est de produire des sons de parole à partir d'une représentation phonétique du message. C'est extrêmement difficile, d'autant plus que le texte à convertir ne contient pas toutes les informations requises pour le transformer en parole, en particulier, les informations prosodiques nécessaires pour rendre naturelle la parole synthétique, sont insuffisamment indiquées.

Dans cette partie du chapitre, nous présentons le schéma général d'un système TTS et nous décrivons ses différentes composantes. En particulier, les différentes méthodes de

synthèse existantes, qui représentent le moyen mis en œuvre pour passer de la représentation symbolique du texte vers le signal acoustique, et les techniques de synthèse (pour la modification de la prosodie et la concaténation des unités de synthèse) sont discutées en détail.

Tableau 1.1 : Consonnes Arabes et leurs notations phonétiques SAMPA

		Bilabiale	Labiodentale	Interdentale	Alvéolaire dentale	Palatale	Vélaire	Uvulaire	Pharyngale	Laryngale
Occlusives	Voisé	ب/[b]			ض/[d.] د/[d]	ج/[Z]				
	Non Voisé				ط/[t.] ت/[t]		ك/[k]	ق/[q]		ء/أ/[ʔ]
Fricatives	Voisé			ظ/[z.] ذ/[D]	ز/[z]			غ/[G]	ع/[H]	
	Non Voisé		ف/[f]	ث/[T]	ص/[s.] س/[s]	ش/[S]	خ/[x]		ح/[X]	هـ/[h]
Nasales	Voisé	م/[m]			ن/[n]					
Vibrante	Voisé				ر/[r]					
Liquide	Voisé				ل/[l]					
Semi Voyelles	Voisé	و/[w]				ي/[y]				

Tableau 1.2 : Voyelles Arabes et leurs notations phonétiques SAMPA

Symbole phonétique	**Symbole Arabe**	**Articulation approximative**
[a]	ـَ	Voyelle brève non arrondie ouverte médiane
[aa]	ـَا	Voyelle longue non arrondie ouverte médiane
[i]	ـِ	Voyelle brève non arrondie fermée antérieure
[ii]	ـِي	Voyelle longue non arrondie fermée antérieure
[u]	ـُ	Voyelle brève arrondie fermée postérieure
[uu]	ـُو	Voyelle longue arrondie fermée postérieure

1.3.1. Définition d'un système TTS

Un système de synthèse à partir du texte désigne l'ensemble des traitements permettant à une machine de transformer un texte écrit quelconque en parole intelligible et naturelle. Il est en effet possible de produire automatiquement de la parole en concaténant simplement des mots ou des parties de phrases préalablement enregistrés. Dans ce cas il est clair que le vocabulaire utilisé reste très limité et que les phrases doivent respecter une structure fixe, afin de maintenir dans des limites raisonnables la quantité de mémoire nécessaire à stocker les éléments vocaux de base. Nous citons à titre d'exemples : l'horloge parlante, ou l'annonce des arrivées dans les stations de train, etc.

Dans le contexte de la synthèse TTS, il est impossible (et heureusement inutile) d'enregistrer et de stocker tous les mots de la langue étudiée. Il est ainsi plus approprié de définir la synthèse TTS comme la production automatique de la parole d'un texte à vocabulaire illimité à partir de sa représentation phonétique [27].

1.3.2. Architecture d'un système TTS

Un système **TTS** comprend un module de Traitement du Langage Naturel ou **NLP** (**N**atural **L**anguage **P**rocessing), capable de produire une transcription phonétique correcte du texte à lire munie d'une description prosodique, et un module de traitement numérique du signal ou **DSP** (**D**igital **S**ignal **P**rocessing), qui transforme l'information symbolique, qu'il reçoit du module NLP, en parole.

La figure 1.2 montre le schéma général d'un système de synthèse TTS. Les différentes méthodes de génération du signal vocal seront décrites dans la section suivante. Alors que les différentes opérations (dites de haut niveau) effectuées dans le module NLP sont examinées brièvement dans cette section.

1.3.2.1. Prétraitement du texte

La première étape effectue le découpage du texte en phrases structurées en mots, puis la réécriture en toutes lettres des nombres, des dates, des abréviations, ainsi que l'identification et la prononciation des sigles. Dans le but de simplifier la synthèse des mots dont la prononciation est irrégulière, un dictionnaire d'exceptions est utilisé pour remplacer les mots irréguliers par leur transcription phonétique.

1.3.2.2. Prononciations des séquences de mots

Une fois la séquence des mots est spécifiée par la procédure de prétraitement, l'étape suivante consiste à déterminer leurs prononciations. Les approches basées sur des règles de transcription sont largement utilisées pour réaliser cette tâche. Elles permettent d'associer un phonème à une ou plusieurs lettres successives en tenant compte des lettres présentes dans le contexte [28]. D'autres approches utilisant les bases de données pour la transcription phonétique ont été également développées. Nous pouvons citer comme exemples : les approches basées sur les arbres de décision statistiques [29] et les approches utilisant les réseaux de neurones entraînés sur de larges dictionnaires de prononciation afin de modéliser la transformation entre les séquences de lettres et les séquences des phonèmes [30].

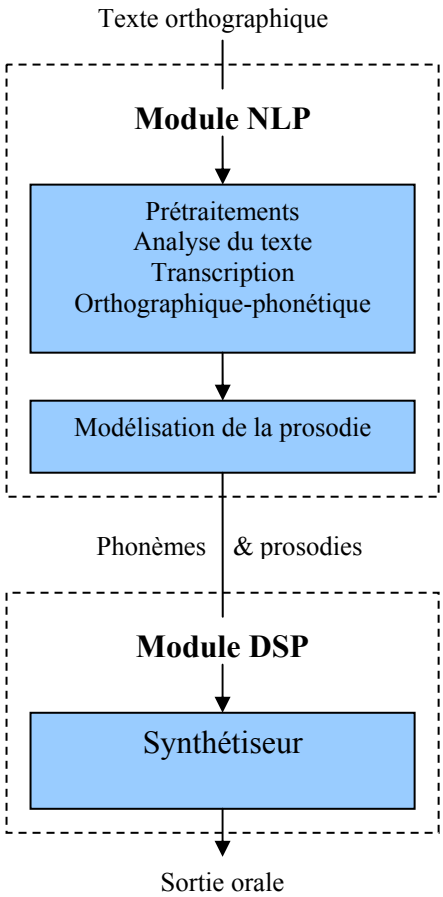


Figure 1.2 : Schéma général d'un système TTS

Pour la langue Arabe, comme certaines autres langues, notamment la langue Espagnole et la langue Finlandaise, il y a toujours une relation étroite entre l'écriture et la prononciation. Ainsi, il est possible de formuler un nombre raisonnable de règles de transcription. Cependant, il existe certaines ambiguïtés et difficultés dans la transcription

graphème-phonème de l'Arabe [10], [25]. Pour pallier à ces difficultés, Y. EL-IMAM [25] et S. BALOUL [10] utilisent une centaine de règles et un lexique d'exceptions dans le module TOP (Transcription Orthographique-Phonétique) de leur système TTS. Nous citons également les travaux de Z. ZEMIRLI [31], et A. SAROH [32] qui se basent sur une analyse morphologique et syntaxique des mots.

Tous ces travaux considèrent le cas du texte arabe voyellé. Cependant, les symboles diactriques sont absents à l'écrit dans la majorité des textes arabes ce qui peut engendrer des ambiguïtés de prononciation dans un système TTS. Pour résoudre ce problème certains travaux procèdent à une vocalisation ou voyellation automatique qui consiste à insérer les symboles diactriques dans le texte avant la synthèse. Une méthode pour la voyellation de la langue persane (qui est très semblable à la langue arabe) a été proposée par F. HENDESSI et al. [33]. Elle consiste à utiliser un réseau de neurones MLP (Multi-Layer Perceptron) à deux couches pour vocaliser chaque mot du texte à synthétiser. Ceci est basé sur le travail de T. SEJNOWSKI et C. ROSENBERG : NetTalk [29]. Au moment où NetTalk permet de transformer un texte écrit en Anglais en une suite de phonèmes correspondant à sa lecture, le réseau de neurones proposé par F. HENDESSI permet de déterminer le type de voyelles de chaque consonne du mot à vocaliser (Figure 1.3).

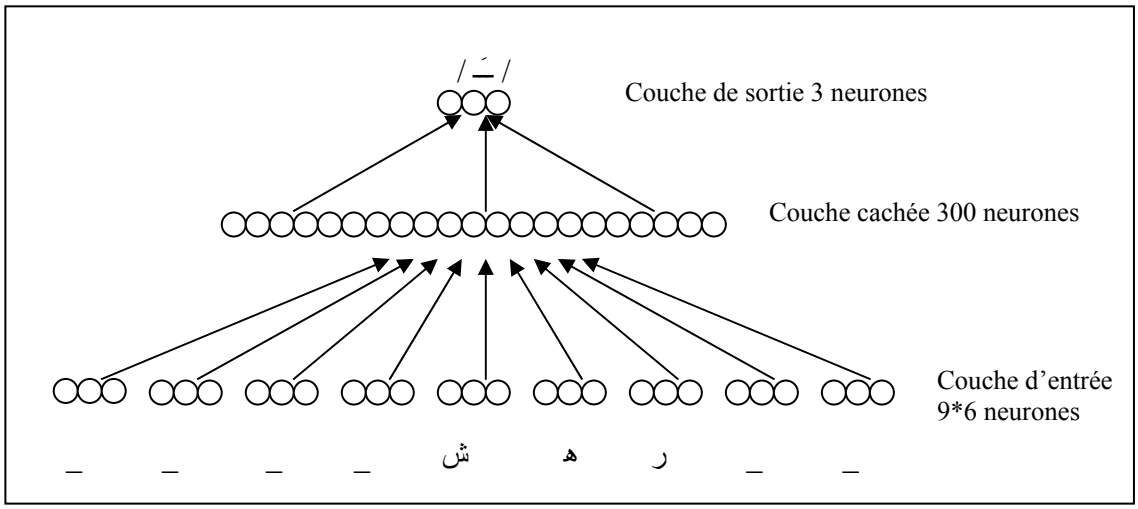


Figure 1.3 : Architecture du réseau neuronal pour la voyellation des mots
Exemple du mot «شهر »

La couche d'entrée comprend 9 groupes de 6 neurones. Chaque groupe correspond à un caractère codé sur 6 bits. Les 9 caractères en entrée forment un contexte local de quatre caractères entourant de part et d'autre un caractère central. La couche cachée contient 300 neurones, alors que la couche de sortie comprend 3 neurones servant à coder la voyelle de la consonne centrale de la fenêtre glissante dans la couche d'entrée. Un modèle de Markov

Caché Ergodique Lisse SEHMM (Smooth Ergodic Hidden Markov Model) est utilisé comme post-traitement pour compenser les éventuelles erreurs générées par le réseau de neurones [33].

1.3.2.3. Génération de la prosodie

Le générateur de prosodie calcule la durée, la fréquence fondamentale, ou le pitch de chaque unité de son afin de pouvoir produire une parole synthétique la plus naturelle possible. Différentes méthodes de modélisation de la prosodie sont discutées dans le chapitre 2.

1.3.3. Méthodes de synthèse de la parole

Une fois l'information concernant les segments de parole à synthétiser et le modèle de prosodie sont générés, l'étape finale du système TTS est de produire le signal de parole synthétique. Les approches principales de cette tâche dépendent du type de la modélisation utilisée. Cela peut être un modèle du système de production vocal humain (synthèse articulatoire), un modèle du signal de parole résultant (synthèse à formants à base de règles), ou bien l'utilisation de segments de parole pré-enregistrés extraits d'une base de données et juxtaposés (synthèse par concaténation).

1.3.3.1. Synthèse articulatoire

Pour produire de la parole, l'être humain met en mouvement ses organes phonatoires (poumons et cordes vocales) et les articulateurs qui modèlent la forme de son conduit vocal (mandibule, langue, lèvres et velum). La génération d'un signal acoustique de parole perceptible nécessite donc une coordination complexe et précise des différents organes, dans l'espace et dans le temps, et implique le recrutement de plusieurs muscles.

La Modélisation des organes réels de la parole est une approche attrayante, puisqu'elle peut être considérée comme étant une modélisation du niveau fondamental du système de production de la parole.

Un synthétiseur articulatoire, comme d'autres modèles source-filtre de production de la parole, consiste en trois éléments : une source d'excitation glottale, un modèle des propriétés acoustiques du conduit vocal et un modèle des effets du rayonnement de l'air au niveau des deux cavités labiale et nasale. Un modèle des plis vocaux à deux ou plusieurs masses, comme celui montré sur la figure 1.4, est souvent utilisé pour fournir le signal

d'excitation glottale pour les sons voisés. Dans ce modèle le mouvement des cordes vocales est simulé par les masses couplées par des ressorts [34,35].

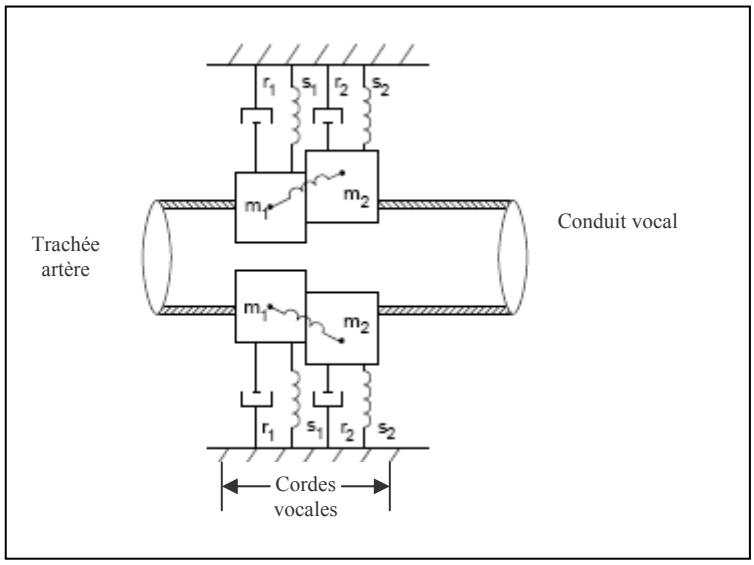


Figure 1.4 : Modèle articulatoire des plis vocaux à 2 masses [35]

En utilisant des données des rayons X prises pendant la production de la parole, des sections transversales du conduit vocal sont obtenues pour les différents sons de la parole. Celles-ci sont utilisées pour construire des modèles statistiques du conduit vocal, tel que celui produit par S. MAEDA [36]. Des valeurs moyennes pour chacun des paramètres suivants sont déterminées pour chaque phonème : la saillie des lèvres, leur taille, la mâchoire, le corps de langue, son dorsum, son apex, la taille du larynx (Figure 1.5). Une troisième dimension est ajoutée à cette section transversale pour créer une fonction d'aire. La transformation de la section transversale en fonction d'aire est réalisée par le modèle de P. PERRIER et al. [37]. Ce modèle repose sur l'analyse d'un moulage du conduit vocal d'un cadavre.

Le rayonnement de l'air au niveau des deux cavités labiales et nasales est souvent modélisé comme un orifice circulaire dans un plan infini. Un modèle plus précis, mais plus difficile à implémenter, basé sur un orifice circulaire dans une sphère, a été également proposé [38].

Même si l'idée d'imiter l'appareil vocal paraît séduisante, Il est évident d'après ce qui précède que la synthèse articulatoire réaliste est un processus extrêmement complexe, et qu'il n'est pas du tout facile de rassembler les données nécessaires, en plus du danger de l'exposition prolongée devant les rayons X. Les recherches dans ce domaine n'ont pas

toujours abouti et, de ce fait, la conception de synthétiseurs articulatoires reste du domaine de la recherche fondamentale.

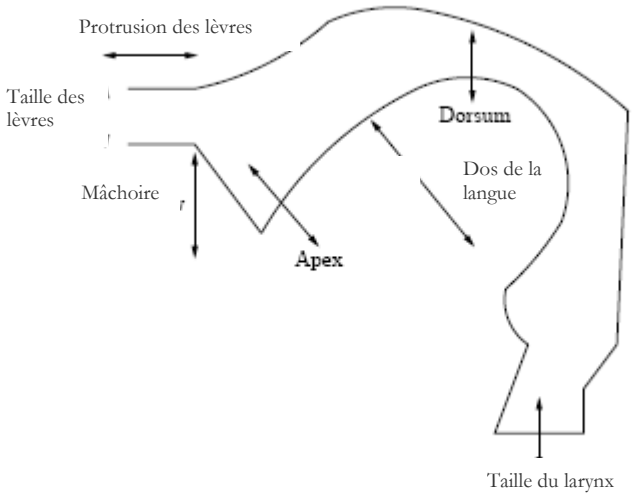


Figure 1.5 : Les 7 paramètres utilisés dans le modèle du conduit vocal de S. MAEDA, montrant les directions des mouvements possibles pour les différents sons

1.3.3.2 Synthèse par règles formantiques

La synthèse formantique à base de règles est une méthode qui a connu beaucoup de succès durant ces dernières décennies. Un ensemble de règles est utilisé afin de déterminer les paramètres nécessaires pour générer de la parole en utilisant un synthétiseur source-filtre à formants. Ces derniers consistent en un ensemble de filtres passe-bande contrôlables, connectés en série ou en parallèle, chacun d'entre eux est utilisé pour générer un seul formant. Des filtres fixes sont aussi incorporés pour modéliser les formants d'ordre supérieur, l'impulsion glottale et les caractéristiques de radiation. Le synthétiseur de J. N. HOLMES [39] est un bon exemple de synthétiseurs à formants parallèle (Figure 1.6). Il exige 11 paramètres de contrôle, adaptés toutes les 10 ms : la fréquence et l'amplitude des trois premiers formants, l'antiformant, l'amplitude du quatrième formant fixe, le pitch et la décision Voisé/Non Voisé [40].

La configuration en cascade ou série est souvent considérée plus attrayante car les amplitudes relatives des formants, dans le cas des voyelles, n'ont pas besoin de contrôle individuel pour chaque formant [41]. Cette configuration donne de bons résultats dans le cas des sons oraux voisés, et elle est facile à implémenter.

Cependant, la configuration parallèle montre des performances supérieures dans la synthèse des sons plosives, fricatifs et nasaux. Pour cette raison, le synthétiseur à formants de D. KLATT [42] contient à la fois des bancs parallèles et série (Figure 1.7).

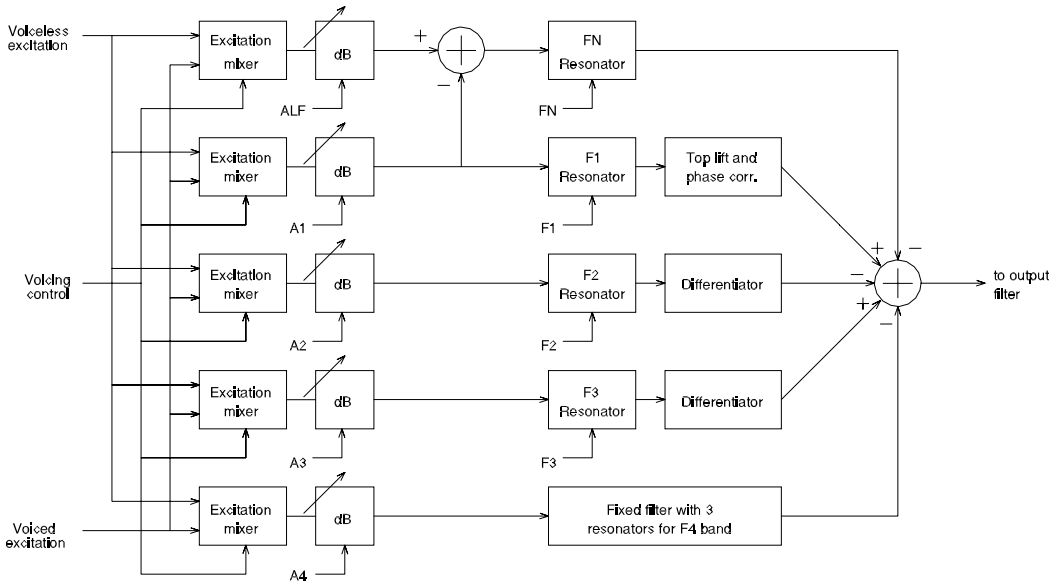


Figure 1.6 : Synthétiseur à formants de J. N. HOLMES [39]

Les synthétiseurs par règles sont basés sur l'idée que, si un phonéticien expérimenté est capable de lire un spectrogramme, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel pour une suite de phonèmes donnés. Une fois le spectrogramme obtenu, il ne reste plus alors qu'à générer l'audiogramme correspondant [27].

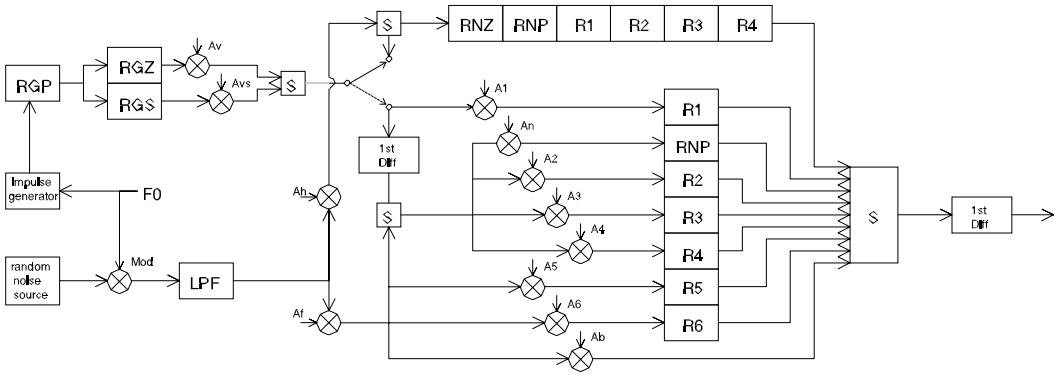


Figure 1.7 : Synthétiseur à formants série-parallèle de D. KLATT(d'après [42])

Dans un premier temps, on fait lire par un locuteur professionnel un grand nombre de mots, généralement de type Consonne-Voyelle-Consonne [CVC] et on les enregistre sous forme numérique. Les mots sont choisis de façon à constituer un corpus représentatif des transitions phonétiques et des phénomènes de coarticulation (l'influence d'un son sur son voisin) dont on veut rendre compte. Ces données numériques sont analysées à l'aide d'un modèle source-filtre à formants , qui a pour rôle de séparer les contributions respectives de la source glottique et du conduit vocal et de présenter cette dernière sous forme compacte, plus propice à l'établissement des règles. On commence par inspecter globalement l'ensemble des données, de façon à établir la forme générale des règles à produire. On précise alors les valeurs numériques des paramètres intervenant dans ces règles (les fréquences des formants, ou les durées des transitions, par exemple) par un examen minutieux du corpus. La mise au point du synthétiseur s'achève par un long processus d'essais-erreurs, afin d'optimiser la qualité de la synthèse.

Lorsqu'un nombre suffisant de règles est établi, la synthèse proprement dite peut commencer. Les entrées phonétiques du synthétiseur déclenchent l'application des règles, qui produisent elles-mêmes un flux de paramètres liés au modèle de parole utilisé. Cette séquence temporelle de paramètres est alors transformée en parole par un synthétiseur à formants, qui implémente les équations du modèle.

La synthèse à formants à base de règles est à la base de quelques systèmes TTS de très haute qualité, tels que le système MITALK [43], le système JSRU [39], INFOVOX SA-101 [44], PROSE-2000 [45] et KLATTALK [46]. Ce dernier a été autorisé au Digital Equipment Corporation pour devenir DECTALK [47].

Les avantages et inconvénients de cette méthode apparaissent de manière évidente : Elle nécessite très peu de mémoire pour stocker les données (seulement les valeurs cibles). Cependant, l'établissement des règles est long et fastidieux. Le nombre des règles peut devenir très grand si l'on souhaite une reproduction fidèle des caractéristiques de la parole humaine. De plus, ces règles dépendent de la langue étudiée. La majeure partie du travail est à refaire si l'on change de langue, et dans une moindre mesure, de locuteur.

1.3.3.3. Synthèse par concaténation

La concaténation d'unités naturelles pré-enregistrées est probablement le moyen le plus simple pour produire une parole synthétique intelligible et naturelle. Les synthétiseurs par concaténation ont une connaissance très limitée du signal qu'ils mettent en forme. La

plupart de ces connaissances se trouvent en effet stockée dans les unités de parole mises en oeuvre par le synthétiseur.

L'un des aspects les plus importants dans la synthèse par concaténation est de bien choisir les unités à concaténer. Avec des unités longues, on obtient une parole naturelle, moins de points de concaténation et un bon contrôle de la coarticulation, cependant le nombre d'unités à stocker et par conséquent l'espace mémoire augmente. L'utilisation d'unités plus courtes nécessite moins d'espace mémoire, mais les procédures de collection et d'étiquetage de ces unités deviennent plus difficiles et complexes.

Diverses combinaisons de diphones (un diphone est un segment compris entre les parties stables de deux réalisations phonémiques adjacentes et contient en son centre toute la zone de transition), de demi-syllabes, et de triphones (qui différent des diphones en ceci qu'ils comprennent un phonème central complet) sont en général retenues, dans la mesure où elles enferment assez correctement les phénomènes de coarticulation tout en ne nécessitant qu'un nombre limité d'unités [27]. Dans le cas de phonèmes ne présentant pas de partie stationnaire, on prend soit la partie la plus stable, soit un triphone, ce qui évite de devoir segmenter la partie transitoire.

Lors de la synthèse proprement dite, ces unités doivent être extraites à partir d'une base de données dont la construction est extrêmement importante. Elle doit pouvoir couvrir toutes les unités requises. Pour cela un corpus textuel (liste de mots, de courtes phrases, voire de textes) dans lequel toutes les unités apparaissent au moins une fois, est enregistré sous forme numérique. Après la segmentation et l'étiquetage de ce corpus, le résultat constitue la base de données qui est souvent stockée en utilisant une certaine forme de codage de la parole.

L'utilisation d'un modèle de parole est souvent maintenue pour trois raisons :

- les modèles bien choisis permettent une réduction de la taille des données, et par conséquent une réduction de l'espace mémoire nécessaire pour stocker ces données ;
- la majorité des modèles de parole séparent explicitement les contributions respectives de la source d'excitation et du conduit vocal. Ceci est mis à profit par le synthétiseur pour résoudre indépendamment (et donc plus simplement) les problèmes de modification de la prosodie des unités et leur concaténation ;

- certains modèles séparent explicitement la parole en deux parties (une partie harmonique et une partie bruit). Ce qui permet aussi d'améliorer les problèmes de juxtaposition des unités et la modification de leur prosodie au moment de la synthèse.

En principe, la base de données doit contenir plusieurs représentations de chaque diphone, couvrant les différents effets de la coarticulation qui apparaissent au niveau de la syllabe ou à un niveau plus haut. Dans ce cas (lors de la synthèse proprement dite), une recherche du meilleur diphone est faite de façon à minimiser les différences perceptuelles entre les éléments concaténés. Un diphone particulier est choisi de façon à minimiser un certain critère, tel que la différence spectrale entre les éléments, ou un coût de concaténation d'unités (dans quelle mesure la juxtaposition des segments choisis amène-t-elle des discontinuités ?).

Dans d'autres travaux, les caractéristiques suprasegmentales des sons sont également prises en considération pour leur sélection dans la base de données. Si l'on reprend le cas d'une synthèse par diphones, on retiendra alors un grand nombre de versions de chaque diphone, différant entre elles par leur durée et leur pitch. L'étape de modification de la prosodie qui viendra juste après la sélection des unités s'en trouvera donc considérablement simplifiée (mais non pas totalement éliminée, puisqu'il est en principe impossible d'enregistrer un corpus reprenant toutes les durées et toutes les courbes mélodiques possibles pour chaque unité) [27].

La dernière génération des systèmes TTS utilise la synthèse par sélection dynamique d'unités non uniformes à base de corpus [48-51]. Les unités utilisées sont de taille variable, multireprésentées et de type polyphones extraites à partir d'un grand corpus de parole. Tout d'abord, en permettant l'utilisation lors du processus de synthèse des sous-unités d'un polyphone, elles profitent au maximum de la quantité de signal de parole disponible dans le corpus. Ensuite, la multireprésentation permet d'associer à une unité symbolique plusieurs réalisations acoustiques enregistrées dans des contextes différents. Les unités de taille variable présentent l'avantage de protéger de la concaténation certains phonèmes. Enfin, le lieu de concaténation des polyphones est situé sur la zone stable de la réalisation phonétique.

Lors du processus de synthèse, la sélection des unités se fait par minimisation d'une fonction de coût combinant par exemple un certain nombre de critères tels qu'un critère

visant à minimiser le nombre de points de concaténation et les discontinuités en ces points, un critère prosodique dont l'objectif est la minimisation des différences entre la prosodie des unités sélectionnées et celle des unités à produire, et un critère lié au contexte phonétique.

Après avoir choisi les unités les plus appropriées, la prosodie correcte doit être imposée par un module de modification de la prosodie. Puisque les segments sont en général représentés sous forme paramétrique, cette opération implique typiquement une modification des paramètres associés à la source (d'où l'intérêt des modèles où ces paramètres sont indépendants de ceux du conduit vocal).

A la sortie du module d'adaptation de la prosodie, les discontinuités de pitch possibles entre segments successifs se trouvent implicitement éliminées. Il reste cependant d'éventuelles discontinuités spectrales. Le rôle du module de concaténation est de les éliminer et ceci dans la mesure du possible, par lissage spectral dans le domaine paramétrique. Ici aussi, le choix du modèle utilisé se révèle être d'une importance primordiale. Bien choisi, il permet, par une simple interpolation linéaire de ses coefficients, de réaliser un lissage spectral qui correspond approximativement au passage naturel d'un son à l'autre [27].

Les techniques les plus utilisées pour la modification de la prosodie et la concaténation des unités dans un système de synthèse TTS sont l'algorithme PSOLA (Pitch Synchronous OverLapp and Add), les techniques LPC (Linear Predictive Coding), LP-PSOLA (Linear Predictive PSOLA) [2], HNM (Harmonic plus Noise Model) [52] et MBROLA (Multi-Band Resynthesis pitch synchronous OLA) [27, 53].

La méthode de synthèse par concaténation fournit une parole synthétique de bonne qualité. Elle est utilisée dans un grand nombre de systèmes commerciaux et expérimentaux tels que les systèmes BRITISH TELECOMM'S LAUREATE [54], AT&T NEXT-GEN [55], PROVERBE et HADIFIX [56] pour l'Anglais. France Télécom au CNET a aussi développé des systèmes TTS pour le Français à base de diphones [57], [58]. Pour la langue Arabe, nous citons les travaux de Y. EL-IMAM [25], M. GUERTI [24] et BALOUL [10] pour la réalisation de système TTS à base de diphones, CHENFOUR et al. ont réalisé le système PARADIS à base de di-syllabes (qui s'étale du noyau d'une syllabe jusqu'au noyau de la syllabe suivante) [59], alors que la synthèse par sélection dynamique d'unités

non uniformes est utilisée dans le système industriel de SAKHR et le système expérimental d'IBM pour la synthèse de l'Arabe.

1.3.4. Les techniques de synthèse de la parole

Dans cette section, nous présentons les principales techniques de synthèse de la parole qui permettent la modification de la prosodie et la concaténation des unités acoustiques extraites d'une base de données dans un système TTS.

1.3.4.1. La synthèse par prédiction linéaire

Les méthodes par prédiction linéaire LPC (linear Predictive coding) sont à l'origine conçues pour les systèmes de codage de la parole, mais elles peuvent également être employées dans la synthèse. Ce sont les systèmes les plus utilisés pour effectuer des modifications prosodiques de la parole. Le modèle LPC est un modèle source-filtre qui représente le signal de parole comme la sortie d'un filtre numérique récursif de type tout pôle (La théorie mathématique de cette technique est décrite dans l'annexe B). Ce modèle est illustré sur la figure 1.8. Il est souvent appelé modèle Auto-Régressif (AR) parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

$$x(n) = gu(n) - \sum_{i=1}^{P} a_i x(n-i) \tag{1.1}$$

Où u(n) est le signal d'excitation qui est un train d'impulsions dans le cas d'un son voisé, et un bruit blanc dans le cas d'un son non voisé. *p* est l'ordre de prédiction linéaire du filtre. Les paramètres du modèle AR sont : la période du train d'impulsions (et par conséquent la fréquence fondamentale ou F_0) pour les sons voisés uniquement, la décision Voisée/Non Voisée (V/NV), le gain g et les coefficients a_i du filtre 1/A(Z), appelé filtre de synthèse.

$$A(Z) = 1 + \sum_{i=1}^{P} a_i Z^i \tag{1.2}$$

Dans un système de synthèse par concaténation le modèle LPC permet un codage rapide des unités de concaténation et il se prête particulièrement bien aux étapes de modification de la prosodie. La modification de la durée est réalisée tout simplement en synthétisant plus ou moins d'échantillons avec les mêmes coefficients de prédiction. La modification de l'intonation est tout aussi triviale, puisque F_0 est un paramètre explicite du modèle : il suffit de changer ce paramètre à la valeur imposée en entrée pour produire un signal ayant la fréquence requise. La concaténation peut être produite en lissant les paramètres du filtre de part et d'autre du point de concaténation [27].

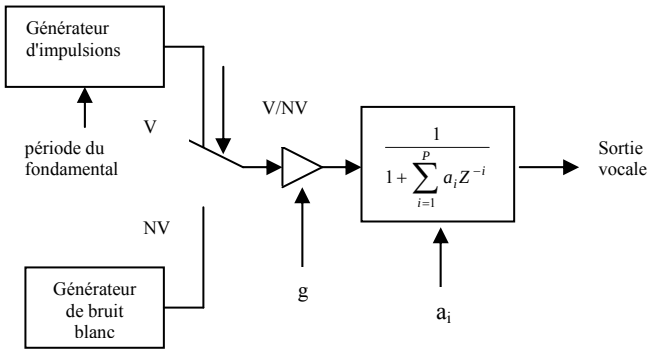


Figure 1.8 : Modèle LPC de production de la parole

Les coefficients de prédiction linéaire a_i ont d'autres représentations : les coefficients LSF (Line Spectral Frequencies), les coefficients de réflexion, les coefficients LAR (Log Area Ratio), etc. Les paramètres LSF sont une variante des coefficients LPC reconnue comme ayant de bonnes propriétés d'interpolation. Si l'interpolation est faite dans le domaine LSF, il est plus facile de garantir la stabilité du filtre de synthèse. Les coefficients LSF sont des paramètres fréquentiels. Généralement les LSF sont concentrés autour des formants. En effet, la proximité de deux coefficients fait apparaître un pic dans le spectre d'amplitude assimilable à un formant (Figure 1.9). A partir des coefficients LSF il est donc possible d'identifier grossièrement les zones auditivement importantes dans le spectre du signal de façon très aisée.

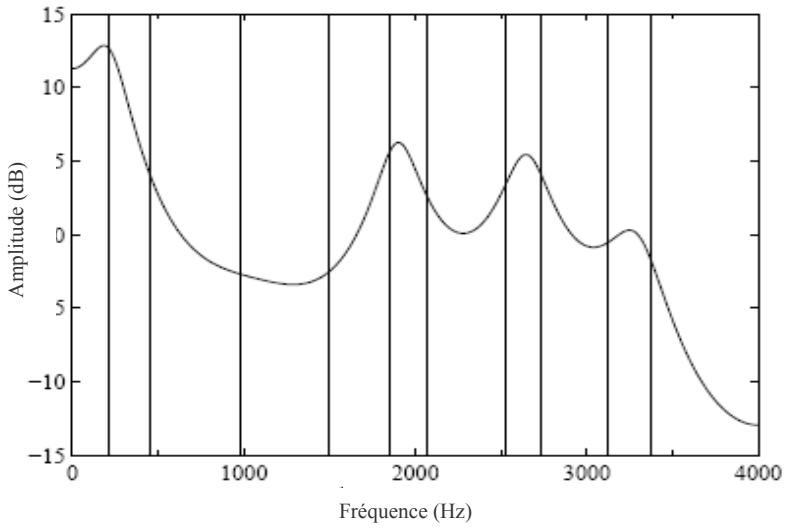


Figure 1.9 : Position des LSF dans le spectre LPC ; les lignes verticales indiquent les positions des coefficients LSF.

Il a été mentionné précédemment que le filtre d'analyse LPC, *A(Z)*, peut être exprimé en termes de coefficient LPC (a_i) comme c'est montré dans l'équation (1.2). Les polynômes symétriques et antisymétriques *P(Z)* et *Q(Z)* d'ordre p+1 peuvent être obtenus à partir de *A(Z)* par :

$$\begin{aligned} P(Z) &= A(Z) + Z^{-(p+1)}A(Z^{-1}) \\ Q(Z) &= A(Z) - Z^{-(p+1)}A(Z^{-1}) \end{aligned} \tag{1.3}$$

où

$$A(Z) = \frac{1}{2}\left[P(Z) + Q(Z)\right] \tag{1.4}$$

Les zéros de *P(Z)* et *Q(Z)* sont entrelacés sur le demi cercle unitaire. Les coefficients LSF sont définis comme étant les positions angulaires de ces zéros qui sont limitées entre 0 et π. Précisément, les coefficients LSF s'écrivent sous la forme :

$$0 = LSF_0 < LSF_1 < LSF_2 < \cdots < LSF_p < LSP_{p+1} = \pi \tag{1.5}$$

LSF_0 et LSF_{p+1} sont toujours égale à 0 et π respectivement. La condition précédente assure la stabilité du filtre de synthèse [60]. Les zéros des polynômes *P(Z)* et *Q(Z)* peuvent être calculés par une méthode qui est bien décrite dans [61], où les polynômes Chebyshev sont utilisés pour trouver les racines

L'inconvénient majeur de la méthode LPC ordinaire est qu'elle représente un modèle tout-pôle, ce qui signifie que les phonèmes qui renferment des antiformants comme les consonnes nasales et les voyelles nasalisées sont mal modélisées. La qualité est également mauvaise avec les courtes plosives à l'intérieur desquelles les événements à l'échelle temporelle peuvent être plus courts que la taille des fenêtres utilisées pour l'analyse. En plus, l'excitation très simplifiée (par rapport au signal résiduel réel) surtout pour les zones voisées de la parole cause aussi une dégradation du signal synthétisé en comparaison avec le signal original.

Pour pallier ces problèmes des modifications et des améliorations ont été appliquées au modèle LPC de base afin d'augmenter la qualité de la synthèse. Parmi ces modifications nous citons la prédiction linéaire modifiée ou WLP (Warped Linear Prediction) qui prend en considération les propriétés humaines d'audition. Ainsi, l'ordre *p* du filtre de synthèse sera réduit significativement [62, 63]. L'idée de base est de remplacer, dans le filtre numérique, les retards unité par les sections passe-tout suivantes :

$$\tilde{Z}^{-1} = D_1(Z) = \frac{Z^{-1} - \lambda}{1 - \lambda Z^{-1}} \tag{1.6}$$

Où λ, -1< λ<1, est un paramètre de modification et $D_1(Z)$ est un élément de retard modifié. Par exemple λ est égale à 0.63 pour une fréquence d'échantillonnage de 22 kHz selon l'échelle de Bark. La méthode WLP fournit une résolution en fréquences meilleure pour les basses fréquences et mauvaise pour les hautes fréquences, ce qui est très similaire aux propriétés d'audition humaine [63].

Une autre modification de la technique LPC, appelée prédiction linéaire multi-impulsionnelle MLPC (Multipulse Linear Prediction Coding) [64], peut résoudre certains des problèmes décrit ci-dessus. La méthode utilise un signal d'excitation complexe formé par un ensemble de plusieurs impulsions par fenêtre d'analyse du signal vocal. La méthode RELP (Residual Excited Linear Prediction) utilise comme signal d'excitation le résiduel ce qui permet une reconstruction exacte du signal vocal. Alors que la méthode CELP (Codebook Excited Linear Prediction) utilise un dictionnaire contenant un nombre fini d'excitations [65].

1.3.4.2. La synthèse par la technique PSOLA

L'algorithme PSOLA a été développé par France Telecom au CNET [66]. La technique ne fait pas la synthèse proprement dite, mais permet une concaténation lisse des segments de parole pré-enregistrés. Elle permet aussi la modification de la durée et du pitch de ces segments.

Plusieurs versions de la technique PSOLA existent. Parmi ces versions, TD-PSOLA (Time Delay PSOLA) en est la plus simple, elle opère directement sur l'onde temporelle. La technique TD-PSOLA repose sur la décomposition du signal temporel échantillonné S(n) en signaux à courts-termes $S_a(n)$ obtenus par multiplication de S(n) par une suite de fenêtres d'analyse $h_a(n)$ de type Hanning. Ces fenêtres sont centrées sur des instants d'analyse t_a.

$$S_a(n) = S(n)h_a(t_a - n) \tag{1.7}$$

Les instants d'analyse t_a, marques de pitch, se succèdent à une cadence synchrone de la fréquence fondamentale sur les segments voisés du signal vocal. Dans les parties non voisées, les marques de pitch sont alors remplacées par un intervalle arbitraire (fixé à 10 ms). La longueur des fenêtres est choisie de façon à ce que deux signaux élémentaires consécutifs présentent un recouvrement mutuel important variant typiquement entre 50% et 75%.

La modification des paramètres prosodiques (durée et fréquence fondamentale) consiste à produire à partir du flux des signaux élémentaires d'analyse un flux de signaux élémentaires de synthèse $S_s(n)$, synchronisés sur une nouvelle suite d'instants t_s, appelés marques de synthèse. Ces modifications correspondent à la duplication ou l'élimination des fenêtres dont l'écartement peut être modifié [67,68] (Figure 1.10 et 1.11).

La synthèse est la dernière étape qui consiste à calculer le signal de synthèse $\hat{S}(n)$ par simple superposition et addition des signaux élémentaires de synthèse qui présentent un taux de recouvrement important de façon analogue aux signaux d'analyse.

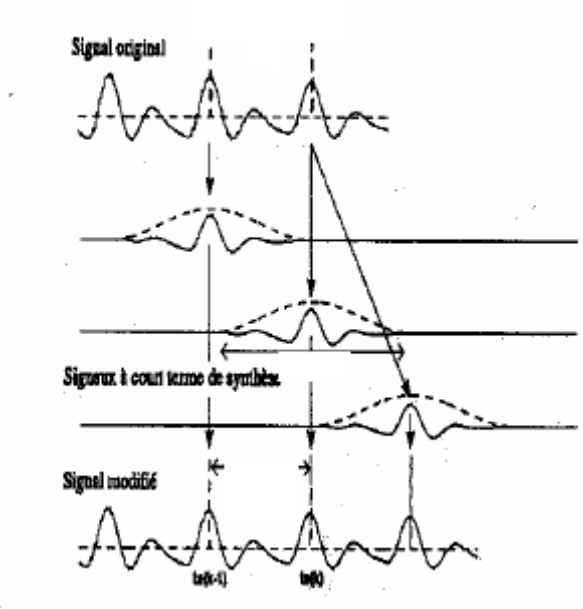


Figure 1.10 : Modification de la durée

La méthode décrite ci-dessus donne une parole synthétique de très bonne qualité par rapport à la synthèse par la méthode LPC. Par sa simplicité, elle peut faire l'objet d'une implémentation en temps réel. Par contre, il est important de noter que la qualité des modifications de la durée et du pitch est très sensible à la position des marques d'analyse.

Une autre variante de PSOLA, FD-PSOLA (Frequency Domain PSOLA) [69], est théoriquement une approche plus appropriée pour les modifications de pitch car elle fournit un contrôle de la fréquence fondamentale indépendamment de l'enveloppe spectrale du signal vocal. Dans cette approche, une enveloppe spectrale globale est obtenue pour chaque signal à court-terme en utilisant, par exemple, les techniques LPC, et une estimation du spectre de la source est obtenue en divisant la Transformée de Fourier Discrète du signal à court-terme par son enveloppe spectrale globale.

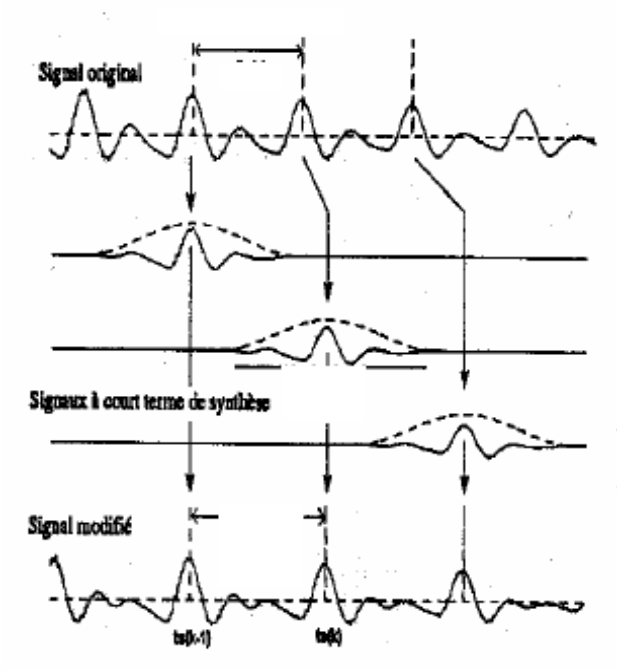


Figure 1.11 : Modification de la fréquence

Le spectre de la source peut être modifié pour imposer la fréquence de pitch désirée. L'enveloppe spectrale peut être aussi modifiée pour améliorer la qualité de la voix, ou bien lisser les frontières des unités à concaténer. Après la modification, les deux spectres seront recombinés et une Transformée de Fourier inverse est appliquée pour générer le signal à court-terme de synthèse. La FD-PSOLA nécessite beaucoup de calcul par rapport à la TD-PSOLA.

Une approche hybride (LP-PSOLA) combinant les deux techniques précédentes LPC et TD-PSOLA tire profit de leurs avantages. Dans cette approche l'algorithme TD-PSOLA est appliqué au résiduel LPC au lieu d'être appliqué au signal vocal lui-même. La LP-PSOLA permet également le lissage de l'enveloppe spectrale aux frontières des unités de concaténation [2].

Une autre version de PSOLA, appelée MBR-PSOLA (Multi-Band Resynthesis PSOLA) a été développée par T. DUTOIT et H. LEICH [70]. Dans cette approche, La base des segments utilisée dans la synthèse est améliorée en utilisant une procédure d'analyse-synthèse MBE (Multi-Band-Excited). Cette procédure permet de re-synthétiser la base de segments de telle sorte que tous les segments soient à pitch constant et à phase fixe (les marques de pitch sont placées au même endroit sur chaque segment plus précisément aux instants de fermeture de la glotte). Ceci élimine le problème de détection des marques de pitch lors de la synthèse et réduit les problèmes de discontinuité qui peuvent apparaître

quand on essaye de concaténer par TD-PSOLA des segments ayant des pitch très différents. Une interpolation spectrale aux frontières des unités à concaténer sera ainsi possible par une simple interpolation linéaire dans le domaine temporel.

1.3.4.3. La synthèse par le modèle harmoniques plus bruit

Le modèle *Harmoniques Plus bruit,* dit modèle HNM (Harmonic plus Noise Model) [71, 72] rassemble divers avantages. Il s'agit d'un modèle hybride en ce sens qu'il décompose les trames de parole en une partie harmonique et une partie bruit. Cette décomposition permet de produire des sons plus naturels. Elle permet en effet la conservation simultanée des énergies basses et hautes fréquences. De plus, les modélisations de l'excitation et du conduit vocal sont ici couplées en un seul système, contrairement à celles des modèles à base de filtres. Enfin, la structure paramétrique du modèle permet d'apporter facilement des modifications prosodiques sur le signal ainsi que de lisser les discontinuités aux points de concaténation. Elle permet en outre un codage de la parole [73]. Le principal défaut de cette méthode reste néanmoins sa charge de calcul importante liée à sa complexité.

En effet, un signal de parole *s(t)* peut être décomposé en une partie harmonique $h(t)$ et une partie bruitée $b(t)$. La partie harmonique modélise la composante quasi-périodique des sons voisés du signal de parole, la partie bruitée modélise la composante aléatoire du signal, c'est-à-dire le bruit de friction et les variations de l'excitation glottique d'une période à l'autre.

$$s(t) = h(t) + b(t) \tag{1.8}$$

avec

$$h(t) = \sum_{l=1}^{L(t)} A_l(t)\cos(2\pi tlF_0(t) + \varphi_l(t)) \tag{1.9}$$

Les paramètres $A_l(t)$, $\varphi_l(t)$ sont l'amplitude et la phase du $l^{\text{ème}}$ harmonique à l'instant t. $F_0(t)$ est la fréquence fondamentale à l'instant t et $L(t)$ est le nombre d'harmoniques inclus dans la partie harmonique à l'instant t. Ces paramètres sont mis à jour à des instants spécifiques t_i appelés instants d'analyse. L'intervalle entre deux instants successifs t_i et t_{i+1} est appelé trame.

Dans le cas des trames voisées, le spectre du signal est divisé en deux bandes limitées par une fréquence variant dans le temps, $F_c(t)$, dite fréquence maximale de voisement ou

fréquence de coupure. Pour les fréquences inférieures à $F_c(t)$, le signal est considéré comme étant purement harmonique et représenté par la partie $h(t)$ (équation 1.9) et au-delà de $F_c(t)$ intervient uniquement une partie aléatoire correspondant au filtrage d'un bruit blanc par le conduit vocal. Pour les trames non voisées apparaît uniquement la partie non déterministe. Le contenu fréquentiel de la partie bruitée est représenté par un modèle AR variant dans le temps. La partie bruitée $b(t)$ peut donc être obtenue en filtrant un bruit blanc gaussien $u(t)$ par un filtre tout pôle $g(t)$ et en multipliant le résultat obtenu par une enveloppe d'énergie $e(t)$.

$$b(t) = e(t)\left[g(t) * u(t)\right] \tag{1.10}$$

Le signal synthétique $\hat{s}(t)$ est simplement obtenu par l'addition de la partie harmonique $h(t)$ et de la partie bruitée $b(t)$:

$$\hat{s}(t) = h(t) + b(t) \tag{1.11}$$

L'implémentation de ce modèle distingue une partie *Analyse* qui paramètre la parole et une partie *Synthèse* qui réalise les modifications prosodiques ainsi que le lissage des discontinuités, avant de générer le signal de parole (voir le chapitre 3).

1.3.5. Autres méthodes et techniques de synthèse

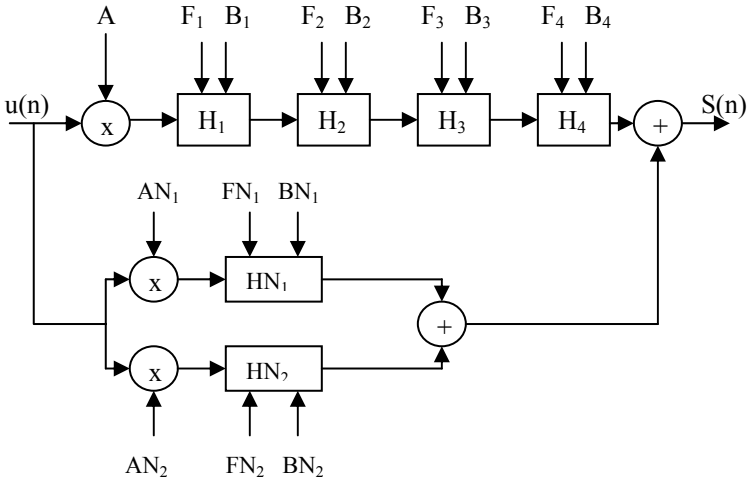
Plusieurs autres approches et expériences ont été également proposées pour améliorer la qualité de la parole synthétique. Certaines de ces approches sont des combinaisons des méthodes de synthèse de base décrites précédemment, car elles montrent différents succès dans la génération de phonèmes individuels. Parmi ces approches, nous citons le système hybride qui combine la synthèse de la parole dans les domaines temporel et fréquentiel [74]. Ce système tire profit des avantages des deux méthodes de synthèse et surmonte certains de leurs inconvénients. Les voyelles et les nasales sont produites exclusivement par la partie paramétrique du système synthétiseur à formants. D'autre part, les consonnes non voisées sont générées dans le domaine temporel. Pour synthétiser les fricatives et les plosives voisées et les transitions voisée/non voisées, les deux composantes du système fonctionnent simultanément.

La figure 1.12 montre le synthétiseur à formants proposé par G. FRIES pour la partie paramétrique du système. Quatre filtres à formants H1,…, H4 en cascade modélisent la production des voyelles. Deux résonateurs parallèles HN_1 et HN_2 sont ajoutés pour réaliser les nasales et les sons nasalisés de plus haute qualité [74].

La composante temporelle du système hybride fournit les formes d'ondes des unités de parole stockées dans un dictionnaire et modifie leur prosodie. Ce dictionnaire contient des variantes allophoniques de fricatives et plosives non voisées, qui sont extraites de la parole naturelle. Les segments de parole non voisés stockés peuvent être également utilisés pour générer les consonnes voisées. Les deux composantes du système sont appliquées simultanément pour produire les fricatives et les plosives voisées ainsi que les transitions voisée/non voisée, par exemple, la composante spectrale quasi-périodique des plosives voisées est générée dans le domaine fréquentiel, alors que la composante stochastique du signal est fournie dans le domaine temporel.

G. FRIES affirme que cette hybridation a amélioré la qualité de la parole synthétique obtenue. En effet, la qualité des fricatives et des plosives est nettement améliorée, en plus des transitions qui sont devenues plus naturelles et sans distorsions au niveau des frontières des segments [74].

D'autres méthodes se basent sur la concaténation de segments sub-phonétiques qui correspondent aux états d'un Modèle de Markov Caché HMM (Hidden Markov Model) entraîné sur une base de données de parole enregistrée par un seul locuteur [75,76]. Alors que d'autres utilisent les réseaux de neurones [40] et les modèles de Markov cachés [77-79] pour déterminer les paramètres de contrôle d'un synthétiseur vocal paramétrique.



F_i, FN_j : Fréquence des Formants
B_i, BN_j : largeur de bande (Bandwidth) des Formants
A, AN_j : Amplitude des Formants

Figure 1.12 : Synthétiseur à Formants de la partie paramétrique
du système hybride proposé par G. FRIES [74]

1.4 Conclusion

Dans ce chapitre, une vue d'ensemble sur la phonétique de l'Arabe standard et du système de production de la parole a été présentée. Nous avons également discuté quelques aspects de la synthèse TTS. Il est clair que, de nos jours, les systèmes TTS génèrent une parole synthétique de haute qualité, mais nous sommes encore loin de délivrer une synthèse parfaite pour toutes les applications possibles. En effet, les trois principales méthodes de synthèse discutées montrent des inconvénients : la grande complexité de la synthèse articulatoire ; l'effort considérable pour l'établissement des règles de transitions entre les différentes représentations phonétiques dans la synthèse par règles ; la difficulté dans la sélection des meilleurs segments à partir d'une large base de données et l'espace mémoire nécessaire pour le stockage de ces segments dans le cas de la synthèse par concaténation, ainsi que les problèmes d'adaptation de la prosodie sans créer des « artéfacts ». Par conséquent, il y a toujours un besoin de rechercher des techniques de synthèse améliorées qui peuvent générer une parole synthétique naturelle avec la prosodie requise surtout quand il s'agit de la langue Arabe où peu de travaux ont été faits dans ce domaine.

CHAPITRE 2
ETUDE DE LA PROSODIE

2.1. Introduction

Nous avons vu dans le chapitre précédent que la génération de la prosodie est une étape importante dans un système TTS. En effet, la prosodie est nécessaire pour que la parole générée imite au mieux la parole humaine. Une description des paramètres prosodiques et une revue des différents modèles existants pour la génération de ces paramètres sont présentées dans ce chapitre.

2.2. Généralités sur la prosodie

Le terme prosodie recèle des notions différentes selon le point de vue adopté pour son étude. Du point de vue acoustique, la prosodie se définit au moyen des paramètres de la fréquence fondamentale (estimation du son laryngien à un instant donné sur le signal), de la durée (intervalle de temps entre deux instants du signal) et de l'intensité (énergie contenue dans le signal). Du point de vue de la perception de la parole, elle concerne l'étude des phénomènes de l'intonation, du rythme et de l'accentuation (variation de hauteur, de la durée et de l'intensité) permettant de véhiculer de l'information liée au sens de la phrase [8,10].

Ce qui tient au niveau du phonème est qualifié de segmental. Par opposition, la prosodie est un phénomène suprasegmental, c'est-à-dire que son domaine d'application est d'un niveau supérieur au phonème. La prosodie intervient sur des unités qui s'étendent du niveau syllabique, à celui de la phrase. Certains chercheurs pensent que des phénomènes prosodiques du niveau du paragraphe et du texte peuvent être mis en évidence [80].

De nombreux travaux ont souligné le fait que les structures linguistiques entretiennent des liens étroits avec les réalisations prosodiques. Chacun s'accorde, aujourd'hui, à considérer que des approches syntaxiques, sémantiques et pragmatiques sont nécessaires pour comprendre les variations prosodiques dans la parole.

La prosodie est à la fois universelle et spécifique à chaque langue [81]. L'illustration la plus parlante de son universalité est la différence intonative entre les modalités déclaratives et interrogatives. Ce qui est particulier à une langue ce sont les stratégies prosodiques utilisées par les locuteurs de cette langue : un texte peut être prononcé de manières différentes selon les caractéristiques anatomiques du locuteur, son origine régionale, sociale, son état émotif, son tempérament, etc. Cela explique que chaque langue a sa propre prosodie, même si elle partage des propriétés avec d'autres langues.

La prosodie est essentielle à la compréhension et au naturel de la parole, et donc indispensable pour un système de synthèse vocale. Cependant, ce n'est pas le seul aspect de son intérêt pour le domaine de traitement automatique de la parole (T.A.P.) : la reconnaissance vocale, le traitement automatique de la parole spontanée (reconnaître les erreurs et les corrections, par exemple) sont des enjeux de plus en plus importants.

2.3. Paramètres prosodiques

Les principaux paramètres acoustiques qui participent à l'étude prosodique sont au nombre de quatre : la fréquence fondamentale (F_0), la durée, l'intensité et le timbre. Chaque phonème a, acoustiquement, une valeur qui lui est propre pour chaque paramètre. Le phonème [a] est en général plus long que le phonème [i], mais ce dernier a, la plupart du temps, une valeur de fréquence fondamentale qui est plus haute. On parle de valeurs intrinsèques. Mais, on sait maintenant que le contexte segmental (l'influence des phonèmes contigus) de production d'un phonème agit aussi sur les valeurs des différents paramètres acoustiques. On parle alors de valeurs co-intrinsèques.

De plus, si nous avons tendance à séparer, dans les observations, les trois paramètres acoustiques, fréquence fondamentale, durée et intensité, il est reconnu qu'ils ne sont pourtant pas indépendants. Des études ont été faites, aucune n'a pu déterminer précisément et sûrement, les relations prosodiques qu'entretiennent ces paramètres dans les stratégies d'un locuteur [82].

Si nous nous plaçons sur le plan perceptuel, la variation dans le temps de ces paramètres correspond à la perception de la mélodie des phrases, de leur rythme et de l'accentuation.

- la mélodie de la phrase correspond à l'évolution dans le temps de la hauteur ;
- le rythme des phrases est perçu grâce à l'enchaînement des durées des segments ;

- l'accentuation est un phénomène de plus haut niveau, qui consiste à mettre en relief une syllabe.

2.3.1. Fréquence fondamentale ou F_0

Physiologiquement la fréquence fondamentale correspond à la fréquence laryngienne produite par les vibrations des cordes vocales. L'énergie de phonation est fournie par l'air des poumons. On peut distinguer des effets à long terme et des perturbations locales au segment (phonème), cela permet de différencier respectivement ce qui tient de la macroprosodie et de la microprosodie. La microprosodie est le résultat de perturbations liées à la prononciation de certains phonèmes, elle est incontrôlée. Quand nous parlerons de prosodie, nous ne considérons que la macroprosodie.

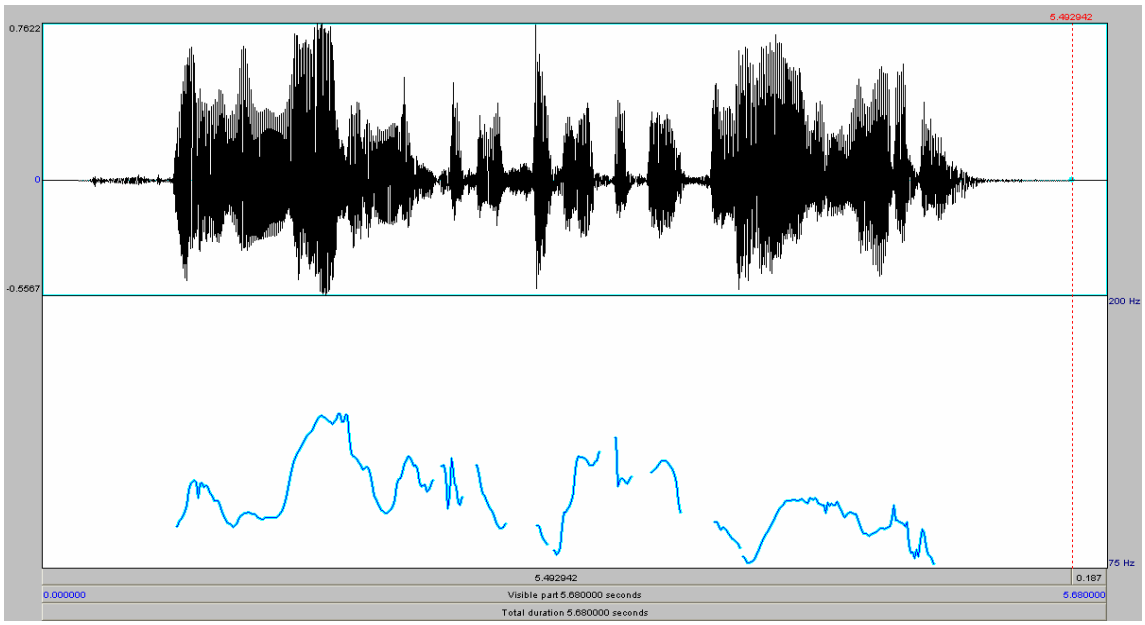


Figure 2.1 : Evolution de la fréquence fondamentale de la phrase arabe :

«من رواد النهضة الحديثة في العالم العربي»

Les variations de F_0 sont les corrélats acoustiques de la mélodie. Du point de vue perceptif, nous parlons de hauteur. Il a été démontré qu'une échelle logarithmique traduit bien mieux qu'une échelle linéaire la perception des hauteurs que nous pouvons avoir. Pour calculer les variations entre deux valeurs de la fréquence fondamentale, nous avons la formule suivante :

$$H = 6 * ech * \frac{\ln(\frac{F_2}{F_1})}{\ln(2)} \qquad (2.1)$$

H est la valeur de la hauteur pour une fréquence F_2 par rapport à une fréquence F_1, dans une échelle logarithmique. Cette échelle est fonction de *ech*. Si *ech* vaut 1, H sera donnée en tons, s'il vaut 2 en demi-ton, 4 en quart de ton, 8 en huitième de ton…

La fréquence fondamentale n'est calculée que sur des parties voisées de la parole, c'est-à-dire principalement les voyelles, et aussi les consonnes voisées. Il existe plusieurs algorithmes pour le calcul de fréquence fondamentale qui peuvent être de type temporel (Autocorrélation, l'AMDF, la cepstrale,…) ou fréquentiel (le peigne spectral, SIFT, Comb, …etc.). La figure 2.1 montre l'évolution de la fréquence fondamentale de la phrase arabe : «من رواد النهضة الحديثة في العالم العربي».

2.3.2. Durée

Parmi les trois paramètres prosodiques, la durée est la plus difficile à préciser, car elle n'est pas directement associable à un corrélat biologique du système phonatoire. Avant de mesurer des durées, il faut cerner correctement les unités à mesurer. On distingue les durées des unités phonétiques : syllabes, phonèmes ou même distance entre voyelles et les durées des pauses.

Comme les autres paramètres, les durées des unités choisies sont largement dépendantes des facteurs de natures linguistiques (accent, position des mots dans les phrases, catégorie grammaticale, etc.) et extra-linguistiques (l'état physique et physiologique du locuteur, débit de parole, expressivité, etc.). Certains d'entre eux peuvent être privilégiés par rapport à d'autres selon le corpus d'analyse et le style de lecture employés [10, 83].

Chaque phonème a une durée intrinsèque et co-intrinsèque. Ces durées sont des caractéristiques des phonèmes. On se rend compte aisément que le phonème [a], pris seul, est plus long que le phonème [b], par exemple.

2.3.3. Intensité

L'intensité est le paramètre le moins étudié [84]. Elle est liée à l'amplitude des vibrations des cordes vocales. L'intensité, associée à l'énergie du signal, est mesurée habituellement en décibels (dB) pour respecter l'échelle perceptive qui est logarithmique. L'énergie contenue dans un segment de signal échantillonné $(s_t)_{t=1,T}$ à support fini est définie par :

$$E_{dB} = 10 * \log(\sum_{t=1}^{T} s_t^2) \quad (2.2)$$

La perception de l'intensité est liée aux variations de fréquence fondamentale [85]. Il existe des corrélations entre F0 et l'intensité. Perceptivement, on peut aussi exprimer l'intensité en sones ou en phons. Ces échelles tentent de rendre compte de la réalité perceptive et de l'interaction des autres paramètres acoustiques.

2.3.4. Timbre

L'enveloppe spectrale d'un signal dans laquelle on observe les formants forme le timbre d'un son. La particularité d'un timbre est due à des propriétés physiologiques, principalement aux formes des cavités de la bouche.

Dans notre étude, nous ne chercherons pas à observer des variations de timbre. Ce paramètre est fortement lié au paramètre d'intensité. Son rôle a été peu étudié sinon pour des conditions particulières de phonation [86] : voix à intensité variable (criée, chuchotée…), réduction de voyelles, etc.

2.4. Intonation

Le terme intonation a deux définitions possibles : au sens strict, ce mot désigne les changements relatifs aux variations de la hauteur de la voix, que certains chercheurs confondent avec le mot mélodie. Le sens le plus étendu de ce terme fait aussi référence aux changements de la durée et de l'intensité. Dans ce dernier cas, il s'apparente au mot prosodie [87]. Nous utiliserons pour la suite de ce document le terme intonation en référence aux variations de la hauteur, dont le corrélat acoustique est le paramètre F_0.

L'étude de la mélodie (intonation) d'une langue est souvent séparée en deux types de phénomènes :

- les phénomènes micromélodiques sont relatifs à des contraintes physiologiques et/ou acoustiques sur l'appareil phonatoire lors de l'articulation de certains types de phonèmes (voir m1 et m2 figure 2.2). Certaines études ont montré que ces phénomènes contribuent à l'identification des phonèmes. Cependant, ils n'ont aucune incidence sur la perception globale de l'intonation. La plupart des modèles de génération de l'intonation existants négligent leur modélisation.
- les phénomènes macromélodiques peuvent eux-mêmes se diviser en deux classes distinctes :

- Les événements de portée locale tiennent compte de tous les événements locaux qui sont relatifs à la mise en relief d'une syllabe ou d'un mot. Ces événements locaux constituent un des indices acoustiques majeurs de la réalisation de l'accent. Ils sont présents au niveau de la courbe mélodique sous forme de pics intonatifs compris chacun, entre deux minima locaux (voir F1, F2, F3, F4 figure 2.2).
- Les événements de portée globale : la seconde concerne les événements mélodiques qui s'étendent sur des portions de parole beaucoup plus longues, voire même sur la totalité de la phrase. Le phénomène mélodique global le plus important concerne la tendance que connaît la fréquence fondamentale à décroître lentement du début à la fin de la phrase. Cette tendance, baptisée "déclinaison" a été remarquée pour plusieurs langues (Hollandaise, Française, Anglaise, Italienne, Japonaise, etc.), notamment dans la langue Arabe [10,88], et elle pourrait bien être universelle. La déclinaison est souvent accompagnée du phénomène de "remise à zéro" ("resetting"), notamment lorsque l'on considère des phrases longues. Ces réinitialisations de la déclinaison seront généralement situées à des frontières de groupes syntaxiques, accompagnées la plupart du temps d'une pause.

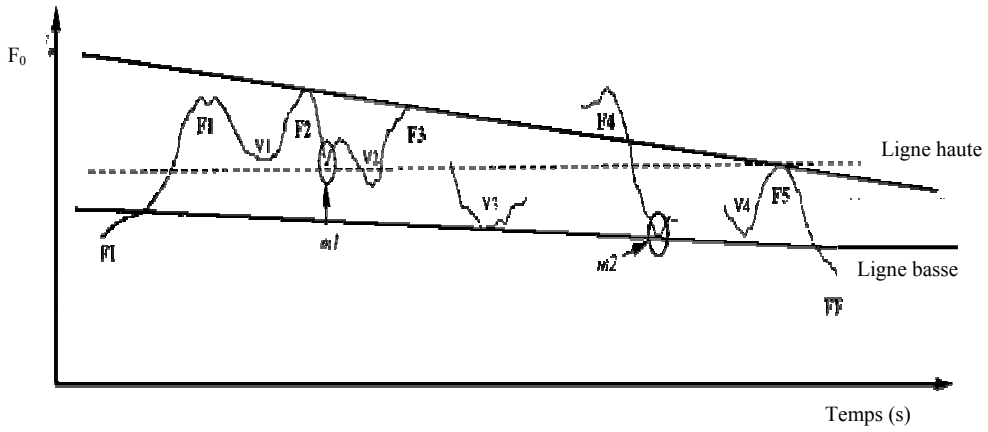


Figure 2.2 : Présentation des principaux paramètres permettant de caractériser les événements mélodiques présents lors d'une analyse acoustique. Avec : FI = Fréquence Initiale, FF = Fréquence Finale, Vx = vallées, Fx =pics mélodiques, mx = creux micromélodiques [87].

BEAUGENDRE [87] a défini deux lignes de déclinaison (une ligne haute et une ligne basse). Ces lignes délimitent le registre du locuteur, c'est-à-dire, la bande à l'intérieur de laquelle se propage la fréquence du fondamental (cf. Figure 2.2).

2.5. Accent

L'accent est un phénomène observable au niveau du mot, on parle de phénomène local. Il occupe une place importante dans la description de la prosodie des langues. L'accentuation est la mise en relief d'une syllabe par rapport à celles qui l'entourent. La perception des accents dans une phrase ne dépend pas de la variation locale d'un paramètre unique. Des expériences montrent que la hauteur, l'intensité et la durée interagissent pour donner, au niveau perceptif, la sensation d'accent.

On distingue généralement deux catégories de langues :

- les langues à accent fixe pour lesquelles la position accentuable du mot est toujours placée sur une syllabe déterminée ;
- les langues à accent libre pour lesquelles la position accentuable dépend de la fonction lexicale et de la structure morphologique du mot.

Dans la langue Arabe, les grammairiens ont ignoré complètement les propriétés accentuelles du langage. Les études qui ont été faites dans le domaine de synthèse de la parole n'ont cessé de confirmer le besoin d'une étude sur l'accent lexical pour la génération d'une prosodie adéquate.

2.5.1. Système syllabique de l'Arabe standard

La syllabe dans la langue arabe obéit à deux règles :

- le noyau syllabique est une voyelle ;
- deux consonnes ne peuvent se suivre sauf enfin de mots et devant une pause.

Les six syllabes considérées sont : [CV, CVC, CVV, CVVC, CVCC, CVVCC] (C≡Consonne, V≡Voyelle). On note ici que les quatre premières syllabes peuvent apparaître au début, milieu ou à la fin du mot. La syllabe la plus fréquente est la [CV] contrairement à la dernière [CVVCC] qui est rare en Arabe standard.

Les syllabes peuvent être classées en deux catégories ; la première syllabe est courte et les autres sont longues. On distingue aussi les syllabes fermées [CVC, CVVC, CVCC,

CVVCC] et les syllabes ouvertes [CV, CVV]. Dans la littérature on peut considérer les syllabes [CVVC, CVCC, CVVCC] comme étant des syllabes sur-lourdes.

2.5.2. Accent lexical de l'Arabe standard

La langue Arabe est une langue à accent variable. Même si les linguistes n'ont pas consacré des études approfondies à l'accent lexical en Arabe, nous pouvons trouver dans la littérature plus d'un algorithme d'accentuation. Cela peut être expliqué par la diversité et l'influence des dialectes Arabes. Selon D. KOULOUGHLI, l'accent est limité aux trois dernières syllabes du mot. Les règles proposées sont [89] :

1. si la dernière syllabe du mot est sur-lourde, celle-ci reçoit l'accent lexical ;
2. si (1) ne s'applique pas et si la pénultième est une syllabe lourde, elle reçoit l'accent lexical ;
3. si (1) et (2) ne s'appliquent pas, l'antépénultième reçoit l'accent lexical.

EL-ANI a proposé un algorithme différent qui considère l'existence de trois niveaux d'accent : l'Accent Primaire (AP), l'Accent Secondaire (AS), et l'Accent Tertiaire (AT) ou niveau inaccentué. Leur position est prédictible : elle dépend du nombre et du type des syllabes dans le mot. Les règles qui permettent de déterminer l'accent dans un mot sont les suivantes [23]:

- si le mot est constitué uniquement de syllabes de type [CV], la première syllabe porte alors l'accent primaire, et les autres sont inaccentuées ;
- si le mot contient une seule syllabe longue, elle porte alors l'accent primaire, et les autres syllabes sont inaccentuées. Les syllabes longues en fin de mot sont ignorées ;
- si le mot est constitué de deux syllabes longues ou plus, la syllabe la plus proche de la fin du mot porte l'accent primaire, la syllabe longue la plus proche du début du mot porte l'accent secondaire, et les autres syllabes sont inaccentuées. Les syllabes longues en fin de mot sont ignorées.

Dans le mot Arabe, la dernière syllabe est toujours exclue dans le processus d'accentuation, et cela quels que soient son type et sa nature. En outre, les prépositions monosyllabiques ([fii], [min], [lan], [maa]) reçoivent un accent secondaire au lieu d'un accent primaire. Dans le cas d'une liaison phonologique entre deux mots successifs, la

dernière syllabe du premier mot, qui possède un accent tertiaire si le mot est isolé, reçoit un accent secondaire [88].

2.6. Rythme

Le rythme est perçu dans la répétition d'un ou plusieurs événements semblables à des intervalles de temps réguliers, c'est la configuration de mouvements ordonnés dans la durée [82,90]. La notion d'un rythme est associée aux variations à long terme de durée [91].

La notion de débit est aussi largement employée dans les études sur le rythme, le débit étant le nombre de syllabes prononcées par unités de temps. Mais d'autres notions importantes existent, au niveau du mot notamment. Ainsi, il ne faut pas confondre la notion de vitesse d'élocution qui est le nombre de mots divisé par le temps de parole, et celle de la vitesse de phonation (ou vitesse d'articulation) qui est égale au nombre d'unités phonétisées divisées pas le temps de parole moins le temps des pauses. Résumons ces définitions par trois formules :

$$Débit = \frac{NbDeSyllabes}{DuréeTempsDeParole} \quad \text{(en syllabes/seconde)} \tag{2.3}$$

$$VitesseElocution = \frac{NbDeMots}{DuréeTempsDeParole} \quad \text{(en mots/seconde)} \tag{2.4}$$

$$VitessePhonation = \frac{NbDeMots}{DuréeTempsDeParole - DuréesPauses} \quad \text{(en mots/seconde)} \tag{2.5}$$

2.7. Modèles de prédiction de la durée segmentale

Les études consacrées à la génération automatique de la durée ont connu beaucoup d'évolution ces dernières années. Nous décrivons, dans ce chapitre, quelques modèles pour la prédiction de la durée segmentale.

2.7.1. Modélisation de la durée par D. KLATT

Les travaux de D. KLATT sont à la base de beaucoup de modèles actuels. IL a proposé un système de règles qui a été implémenté dans le système MITALK [92] en exploitant des informations de la littérature phonétique concernant les facteurs affectant la durée segmentale. La durée de chaque phonème est calculée en utilisant l'équation suivante [83] :

$$DUR = (INHDUR - MINDUR) * \frac{PRCNT}{100} + MINDUR \tag{2.6}$$

Où INHDUR et MINDUR : sont respectivement les durées intrinsèque et minimale du phonème. PRCNT représente un pourcentage déterminé de façon cyclique par application des règles. La valeur finale du paramètre PRCNT est obtenue en multipliant les facteurs qui influent sur la durée segmentale tels que le contexte phonétique et l'environnement syntaxique.

Le modèle de D. KLATT qui était multiplicatif dans ses premières versions (1976), a évolué vers un modèle additif et multiplicatif. Dans ce même contexte, nous pouvons citer d'autres modèles multiplicatif et/ou additif qui prédisent la durée phonémique tels que les modèles de D. O'SHAUGHNESSY [93], le modèle de France Télécoms R&D [94] pour les langues européennes, et les modèles de Z. ZEMIRLI [95] de A. AMROUCHE [96], de S. BALOUL [10], de G. DROUA [97] pour la langue Arabe.

2.7.2. Modèle statistique linéaire : Modèle somme de produits

J. V. SANTEN a développé un modèle linéaire qui est basé sur une collection d'équations déterminées à partir d'informations phonétiques et phonologiques préalables ainsi que des informations rassemblées en analysant des bases de données. Il l'appelle « modèle somme de produits ». L'équation (2.7) montre un modèle somme de produits typique dont les variables sont à calculer manuellement à partir des bases de données par des méthodes des moindres carrés standards [98] :

$$DUR(Voyelle:/e/, Next:Voisé, Pos:Final) = \alpha(/e/) + \delta(Final) + \beta(Voisé) * \gamma(Final) \qquad (2.7)$$

L'équation (2.7) indique que la durée de la voyelle /e/ qui est suivie par une consonne voisée, et se trouvant en position finale d'expression est calculée en prenant la durée intrinsèque de la voyelle α(/e/), et ajouter un certain nombre de millisecondes à cause de sa position en expression finale [δ(finale)], et finalement ajouter l'effet de voisement post-vocalique [β(Voisé)] modulé par [γ(Final)] (un effet indiquant que la voyelle se trouve en position expression finale) [98].

2.7.3. Prédiction des durées des syllabes par réseaux de neurones

Le modèle de W. N. CAMPBELL repose sur l'hypothèse que l'organisation temporelle d'un énoncé se fait à un niveau supérieur sur le plan phonémique. Deux étapes se distinguent, dans la mise en œuvre de ce modèle, la première est la prédiction des durées syllabiques et la seconde est la prédiction à l'intérieur de chaque syllabe des durées phonémiques. Un processus d'apprentissage automatique permet la prédiction des durées

syllabiques. Il emploie les réseaux de neurones pour l'apprentissage parce que l'on suppose que ces derniers peuvent apprendre les interactions fondamentales entre les effets contextuels. Ils devraient pouvoir représenter le comportement régi par des règles qui est implicite dans les données (c'est précisément la raison pour laquelle nous les utilisons pour tous les aspects de la modélisation prosodique dans cette étude). Si les réseaux peuvent coder les interactions fondamentales, ils en feraient de même avec des données non rencontrées auparavant [99].

En ce qui concerne les durées segmentales, leur distribution est donnée par le calcul d'un coefficient d'allongement (déviation par rapport à la moyenne). Il propose que tous les phonèmes d'une même syllabe aient le même facteur d'allongement z : le z-score. Le z-score de chaque réalisation phonémique du corpus d'étude est calculé par :

$$z_{réalisation} = \frac{(durée\,observée_{réalisation} - \mu_{phonème})}{\sigma_{phonème}} \tag{2.8}$$

Où $\mu_{phonème}$ et $\sigma_{phonème}$ sont la moyenne et l'écart type obtenus à partir des durées absolues des réalisations de chaque phonème du corpus. Ainsi, chaque durée d'une réalisation phonémique est normalisée par utilisation du z-score (moyenne=0 et écart-type=1).

Les durées des syllabes étant déterminées par le réseau neuronal, le modèle calcule alors le z-score associé à chaque syllabe en résolvant l'équation ci-dessous :

$$Durée(syllabe) = \sum_{i=1}^{n} \exp(\mu_i + z\sigma_i) \tag{2.9}$$

La somme concerne les éléments phonémiques de la syllabe, z est le z-score associé à cette syllabe et la paire (μ_i et σ_i) contient respectivement la moyenne et l'écart type associés au phonème i et obtenus à partir des logarithmes des durées des réalisations (exprimées en millisecondes) de ce phonème dans le corpus. Ainsi, la durée de chaque phonème de la syllabe est calculée en utilisant l'équation (2.10).

$$durée(phonème\ i) = \exp(z.\sigma_i + \mu_i) \tag{2.10}$$

2.8. Modèles de l'intonation

Dans le cadre de la génération automatique de l'intonation, nous abordons ci-dessous quatre principales méthodes de modélisation, chacune d'entre elles ayant déjà été appliquée à différentes langues.

2.8.1. Modèle de H. FUJISAKI

Cette méthode est fondée sur la présupposition que la courbe mélodique est construite à l'aide de deux types de commandes (commande d'accent et commande de phrase). Elle décrit une commande globale pour la phrase ou le syntagme, sur laquelle se superposent des événements locaux dus aux accents ou aux réalisations de frontières prosodiques. Les fonctions utilisées pour chacune de ces commandes sont d'inspiration physiologique et visent à rendre compte de l'activité musculaire laryngienne mise en jeu lors de la production de la mélodie.

Le modèle de H. FUJISAKI [100], inspirés des travaux de S. OHMAN [101], considère que les contours de F_0 sont le résultat de la superposition de deux types de composantes mélodiques : les composantes de phrases **(**qui correspondent à la réponse à des impulsions d'un système linéaire du second ordre**)** et les composantes d'accents **(**qui correspondent **à** la réponse à des "fonctions échelon" d'un autre système linéaire du second ordre).

2.8.2 Méthode par points cibles (ou approche tonale)

Cette méthode considère que l'information mélodique est contenue dans les extrema de la courbe. Ce modèle est d'inspiration phonologique car il s'agit de simplifier la courbe intonative par une suite de tons discrets (points cibles) reliés entre eux par des fonctions de transition quadratiques [102, 10] ou linéaire [87].

Les points cibles sont généralement localisés sur les parties du signal qui correspondent à un phonème voisé ; ce qui donne pratiquement une cible par syllabe. MOMEL, proposé par D. HIRST et D. ESPESSER [103], par exemple, est un algorithme de modélisation automatique de la courbe de la fréquence fondamentale. Cet algorithme consiste en un lissage par régressions modales sur des valeurs de F_0 éventuellement corrigées. Ensuite à partir de l'estimation de la courbe, il sélectionne des parties représentatives perceptivement. Enfin dans ces ensembles, il ne retient qu'une seule valeur cible par segment. Cela donne approximativement une valeur cible par syllabe (il s'avère nécessaire quelquefois qu'une " correction " manuelle soit nécessaire). L'illustration ci-dessous donne le résultat de l'algorithme MOMEL pour la phrase arabe « من رواد النهضة الحديثة في العالم العربي ».

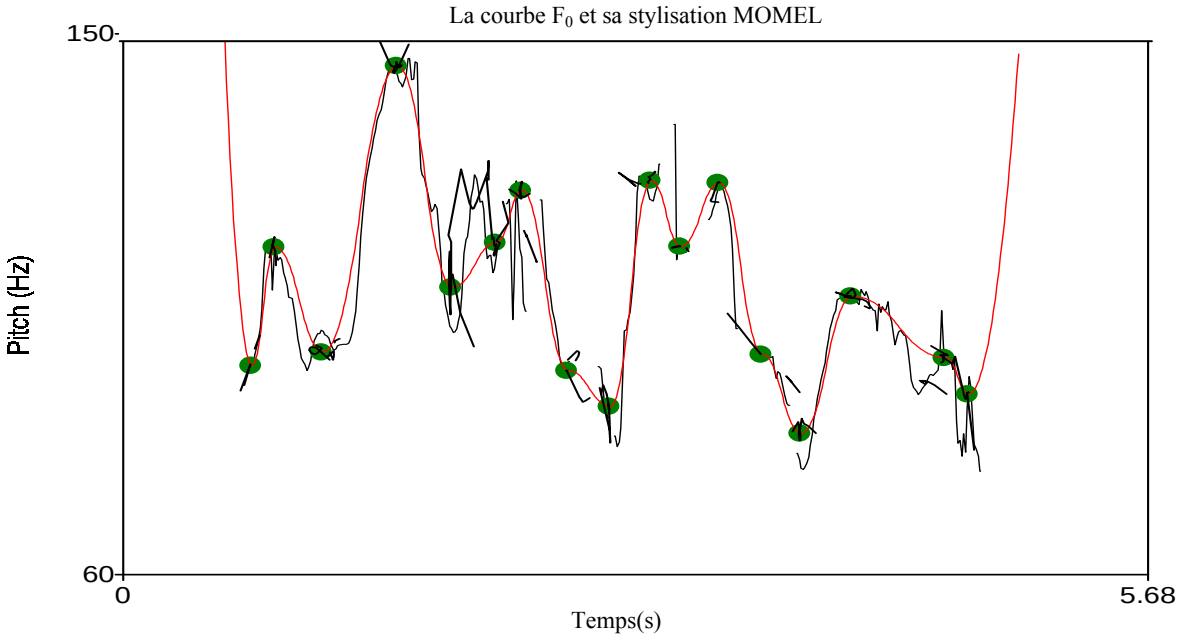


Figure 2.3 : Choix de valeurs cibles par MOMEL de la phrase arabe
« من رواد النهضة الحديثة في العالم العربي ».

Des symboles sont ensuite associés aux différents points cibles, ce qui permet de dériver une représentation formelle à partir de la substance phonétique.

Le plus connu des systèmes de codage symbolique est ToBI (Tones and Break Indices system) [104]. Il s'est profilé comme le système de référence pour la transcription de l'intonation de la langue anglo-américaine. Ce système propose de décrire l'intonation par une succession de tons (hauts/bas) placés sur des cibles spécifiques dans les expressions verbales; un ton est placé notamment sur chaque syllabe accentuée (pitch accents) et à chaque frontière intonative (phrasal tones, final boundary tones). Ce système a été assez largement adopté pour la transcription de l'intonation de l'Anglais et a été adapté à d'autres langues (Exemple pour la description du Coréen : K-ToBI, pour l'Allemand G-ToBI). P. MERTENS a développé un système similaire pour la description de l'intonation du Français [105].

Le système ToBI n'est pas athéorique. Il est basé sur un modèle de séquences tonales (Tone Sequence Model) qui a été originalement proposé par J. PIERREHUMBERT [106] pour l'Anglo-Américain. Ce modèle de génération de l'intonation vise à décrire des règles intonatives fondamentales, ainsi qu'un ensemble de règles de transformation qui seraient capables de générer les séquences de tons réalisées dans différents contextes (pour une synthèse de ce modèle voir [107]).

2.8.3. Modélisation sous forme de contours mélodiques stockés

D’autres chercheurs proposent de modéliser la mélodie à l'aide de contours [87, 108-111]. L'hypothèse de leurs travaux est que la mélodie peut être représentée comme la concaténation de contours ou patrons mélodiques pré-stockés.

Ces patrons sont issus d'observations directes de la courbe de fréquence fondamentale. Pour F. BEAUGENDRE, la variabilité de ce paramètre impose que l'on s'appuie sur des considérations perceptives pour obtenir des simplifications, des stylisations de ces courbes. Les stylisations permettent de limiter le nombre de contours et facilitent ainsi les observations.

Notons toutefois, que ces systèmes visent à découper une phrase en groupes prosodiques généralement liés à la syntaxe à l’aide de règles plus ou moins complexes puis à représenter ensuite l’intonation à l’aide de contours types enregistrés dans un lexique.

Notons également qu’il est possible d’obtenir ces contours par des approches statistiques ou par réseaux de neurones avec en entrée les marqueurs de phrases, syntagmes et groupes prosodiques.

2.8.4. Génération de contours intonatifs à partir de systèmes statistiques ou RN

Parmi les développements les plus récents dans le domaine de la génération automatique de l’intonation, on trouve un intérêt grandissant pour les techniques d’apprentissage automatique telles que les réseaux de neurones [112], les arbres de régression et de classification [113], les modèles de Markov cachés [114], les techniques de régression linéaire [115] ou d’autres méthodes stochastiques [116]. L’ensemble de ces techniques se base sur l’analyse de corpus de parole préalablement étiqueté prosodiquement.

A l'aide de ces techniques, Il s'agit d'associer la description linguistique et la substance prosodique. Ces modèles nécessitent d'utiliser une caractérisation linguistique appropriée et de prélever des valeurs de substance prosodique qui permettent de rendre compte le mieux possible des phénomènes prosodiques pertinents.

C'est surtout dans la préparation des paramètres de caractérisation que réside la plus grande partie du travail. Si ces paramètres sont bien choisis, on peut alors arriver à générer des contours mélodiques de très bonne qualité. L’approche que nous proposons dans ce

travail pour la modélisation de l'intonation se trouve directement dans la lignée de ces méthodes.

2.9. Modèles de l'intonation de la langue Arabe

La synthèse de l'intonation de la langue Arabe a été réalisée pour la première fois avec le développement d'un système TTS basé sur la synthèse par règles [117, 118]. Le modèle intonatif établi est basé sur la concaténation des contours mélodiques des mots [88]. Dans une phrase, chaque contour intonatif d'un mot est généré en utilisant les règles d'accentuation. Le contour du mot consiste en un mouvement croissant et un autre décroissant. Le maximum du contour est corrélé avec l'accent principal. Sur le plan perceptif, la parole de synthèse résultant de ce modèle présente une monotonie indésirable. Cela peut être expliqué par le fait que le modèle garde tous les pics du contour F_0 sur le même niveau intonatif. En plus, la descente brutale de F_0 au niveau de certaines frontières de mots est disruptive de point de vue perceptif [88]. Pour résoudre ces problèmes, A. ZAKI et al ont proposé un modèle se basant sur [88] :

- l'introduction du phénomène de déclinaison dans les règles ;
- le caractère accentué ou non des syllabes, en prenant en considération les trois niveaux d'accents primaire, secondaire et tertiaire ;
- le fait que l'information mélodique est contenue dans les extrema de la courbe intonative ;
- l'utilisation d'un ensemble de règles phonologiques et phonotactiques pour le lissage du contour F_0.

Deux droites de déclinaison sont définies, l'une supérieure et l'autre inférieure, à l'intérieur desquelles la fréquence fondamentale varie. Sur ces droites se trouvent les points cibles de l'intonation.

Les deux lignes de déclinaison sont calculées par la méthode des moindres carrée sur des contours intonatifs d'un corpus de phrases affirmatives. La ligne de base et la ligne haute correspondent à la meilleure droite passant respectivement par les minima et les maxima locaux. Les pentes et les intersections de ces droites avec l'axe des ordonnées sont ensuite calculées en fonction du nombre de syllabes dans une phrase.

Ainsi lors de la synthèse de F_0, connaissant le nombre de syllabes dans une phrase quelconque, une fonction linéaire est utilisée pour le calcul des lignes de déclinaison

$$Y=S(N)*n+I(N) \quad (2.11)$$

Avec N est le nombre de syllabes, n est la position de la syllabe dans la phrase, S(N) et I(N) représentent respectivement la pente et l'intersection avec l'axe des ordonnées.

Ensuite, ils utilisent le système symbolique INTSINT [119] (T, B, M, H, U, D, L) pour définir le pitch d'un point cible. T, B, M sont les pitchs absolus. Ils dépendent des règles d'accentuation. H, L, U et D sont les pitchs relatifs, résultant des règles phonotactiques. T, M, et B sont associés respectivement aux accents principal, secondaire et tertiaire.

Les règles phonotactiques sont :

- quand il y a réalisation de la séquence suivante de points cibles T-B-T, le pitch cible B est remplacé par H.
- Dans la séquence M-B-T, le point cible B est remplacé par U
- Dans la séquence T-B-M, le point cible M est remplacé par D
- Une succession de deux ou plusieurs points cibles B ne sera pas toute alignée avec la ligne basse. Seul le dernier point cible B le sera, les autres sont remplacés par L.

Les symboles absolus B, M et T sont alignés respectivement avec les lignes basse, moyenne et haute du registre du locuteur. Pour les symboles relatifs (H, U, D, L), les valeurs de F_0 sont calculées en tenant compte des informations contextuelles, et du pitch des points cibles à gauche et à droite. Les transitions entre les points cibles sont des fonctions exponentielles [88]. Les règles phonotactiques proposées par A. ZAKI et al. peuvent être illustrées par l'exemple présenté dans le tableau 2.1.

Tableau 2.1 : Exemple de la représentation phonologique de points cibles [88]

Phrase	مستودعاتهم التي في المصنع												
Syllabation	mus	taw	da	Haa	tu	hu	mul	la	tii	fil	mas.	na	Hi
Type de la sykllabe	CVC	CVC	CV	CVV	CV	CV	CVC	CV	CVV	CVC	CVC	CV	CV
Accent	AT	AS	AT	AP	AT	AT	AS	AP	AT	AT	AP	AT	AT
Pitch cible	B	M	U	T	L	B	M	H	L	B	T	L	B

BALOUL a proposé une autre approche pour la modélisation de la courbe mélodique fondée sur les tronçons (groupes prosodiques). Cette approche consiste à diviser la phrase en groupes de mots non récursifs, baptisés *chunk* en Anglais, *tronçons* en Français, sans nécessairement les mettre en relation les uns avec les autres. Les mots appartenant à un

même tronçon se caractérisent par des liens syntaxiques forts : ainsi, leur ordre dans le tronçon est rigide comparé à l'ordre des tronçons dans la phrase, qui est relativement flexible. D'un point de vue prosodique, le tronçon ne peut être scindé ni par une pause ni par une frontière intonative [10].

BALOUL a appelé accent de tronçon, l'accent réalisé comme le pic mélodique du tronçon. Dans un tronçon initial ou final, il correspond à l'accent primaire porté par le premier mot lexical du tronçon ; dans un tronçon intermédiaire, il correspond à l'accent primaire porté par le dernier mot du tronçon.

Théoriquement, sur la courbe de F_0 d'un mot isolé arabe, le maximum de la fréquence fondamentale se situe sur la syllabe qui porte l'accent primaire, et au niveau de la phrase, tout mot arabe garde son accent lexical. Les minima de la courbe mélodique se réalisent le plus souvent sur les dernières syllabes des mots. Le degré d'accentuation des mots diminue au fur et à mesure qu'on se rapproche de la fin d'un tronçon initial ou final, et augmente dans le cas d'un tronçon intermédiaire [10].

Après le découpage de la phrase en tronçons, les points cibles sont déterminés à partir des deux droites de déclinaison et des règles d'accentuation. La ligne haute de déclinaison est la droite qui passe par les maxima mélodiques des tronçons. La ligne de base est la droite qui passe par les minima mélodiques des mots. Enfin, la courbe mélodique est simplifiée sous la forme d'un enchaînement de segments de droites entre les points cibles (Figure 2.4).

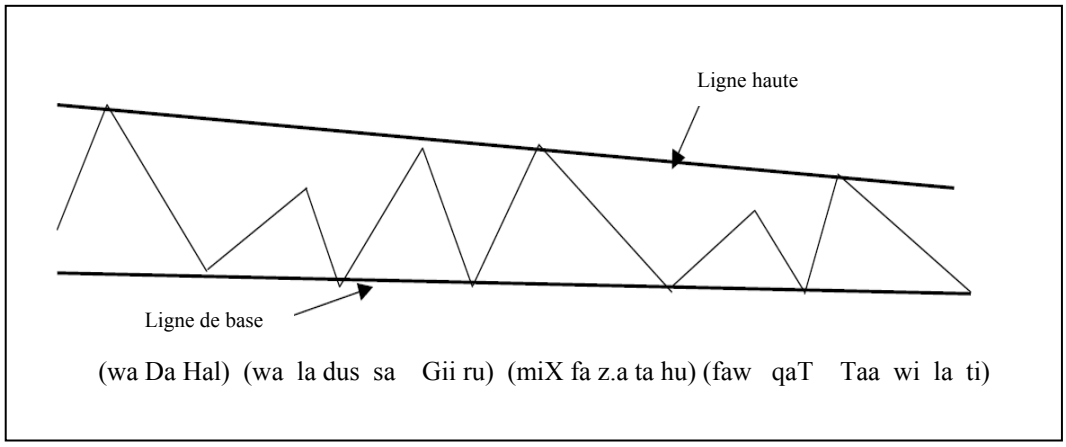


Figure 2.4 : Exemple de contour mélodique d'une phrase déclarative [10]

2.10. Conclusion

Dans ce chapitre nous avons examiné de près quelques modèles existants pour la prédiction de la durée segmentale et la génération des contours intonatifs. Nous avons discuté également des modèles (quoique peu) qui ont été appliqués pour la synthèse de la prosodie Arabe. Ces modèles utilisent des règles qui ne peuvent généralement pas décrire les relations non linéaires entre les informations linguistiques et les paramètres prosodiques. Les règles sont habituellement aussi générales que possibles et des exceptions tendent à augmenter et compliquer l'ensemble de ces règles. En plus de la difficulté d'élaboration de ces règles qui est basé sur l'expertise introspective de différents chercheurs. Tous ces problèmes nous incitent à utiliser les réseaux de neurones pour la génération des différents paramètres prosodiques dans notre système TTS en Arabe standard.

CHAPITRE 3

MODELES DE PAROLE ET MODIFICATION DE LA PROSODIE

3.1. Introduction

Vu que la majorité des systèmes TTS actuels utilisent la méthode de concaténation pour générer le signal de parole, les techniques de modification de la durée et de l'intonation des sons extraits d'une base de données sont extrêmement importantes. Dans ce chapitre, nous examinons l'implémentation de quatre modèles susceptibles d'apporter la haute qualité recherchée, à savoir les modèles LPC, TD-PSOLA, LP-PSOLA, et le modèle HNM.

3.2. Détails d'implémentation des quatre modèles

Dans cette section, nous présentons les détails d'implémentation des quatre modèles que nous avons expliqué dans le chapitre 1.

3.2.1. Modèle auto-régressif LPC

Nous avons mis au point une synthèse LPC classique, avec un ordre de prédiction de 10 (pour une fréquence d'échantillonnage f_S de 8 KHz). La représentation LSF des coefficients LPC est choisie à cause de ses excellentes propriétés d'interpolation. Ces coefficients sont calculés toutes les 10 ms sur des fenêtres de 30 ms. L'adaptation de la prosodie est immédiate (pitch et durée sont des paramètres explicites du modèle).

3.2.2. L'algorithme TD-PSOLA

TD-PSOLA est une technique non paramétrique bien connue en synthèse de la parole pour son faible coût de calcul. Elle repose sur une décomposition du signal temporel en fenêtres recouvrantes synchronisées sur la fréquence de vibrations des cordes vocales (fréquence fondamentale ou pitch ou encore F_0). Cette technique nécessite peu de calculs, mais en revanche la détection de la fréquence fondamentale et la localisation des marques de pitch indiquant les centres des fenêtres s'avèrent obligatoires. De nombreux algorithmes permettent l'extraction de la fréquence fondamentale. Dans notre cas nous avons

implémenté trois méthodes : la méthode d'autocorrélation, la cepstrale et celle du filtrage inverse SIFT (Simple Inverse Filtering Tracking). La technique TD-PSOLA nécessite également, la connaissance des marques de pitch. Ces marques doivent être positionnées à la même place dans toutes les périodes temporelles. Les instants de fermeture glottale ou CGI (Glottal Closure Instants), c'est-à-dire l'instant où le conduit vocal se ferme, sont généralement choisis comme marqueurs de périodes. Plusieurs algorithmes [120-122] existent pour déterminer ces CGI. Dans notre cas, nous nous sommes intéressés à un algorithme de programmation dynamique qui offre une bonne précision sur la localisation des marques de pitch [120].

3.2.3. Modèle LP-PSOLA

LP-PSOLA est une méthode qui combine les deux méthodes LPC et TD-PSOLA et tire profit de leurs avantages. Elle consiste à modifier les caractéristiques prosodiques de l'erreur résiduelle de prédiction linéaire par la méthode TD-PSOLA et ensuite ajouter l'information spectrale par filtrage LPC.

La modification des paramètres prosodiques devient alors une tâche qui consiste à séparer la composante de l'excitation et celle du conduit vocal, ce qui permet un contrôle explicite du spectre de la parole synthétique.

L'avantage de faire les modifications prosodiques au niveau du résiduel à la place du signal lui-même est que les distorsions spectrales dans les fréquences des formants sont plus faibles (le spectre du résiduel est plat) [123,124].

La synthèse LP-PSOLA fonctionne de façon identique avec n'importe quel modèle de codage basé sur la LPC (CELP, MLPC, MELP, etc.). Une approche consiste à ne pas coder du tout le résiduel et utiliser cette erreur de prédiction directement comme entrée du module PSOLA [68], et c'est cette dernière approche qui nous intéresse et qui a été implémenté dans notre travail.

L'implémentation de cette technique passe par les étapes suivantes :

Analyse

A l'aide de la méthode LPC, on détermine pour chaque fenêtre (de 30 ms avec un pas de 10 ms) un filtre tout pôle modélisant l'enveloppe spectrale du signal de parole (nous avons choisi l'ordre du filtre égale à 10 pour fs = 8 kHz). Ensuite, nous obtenons la composante source d'excitation par filtrage inverse du signal d'entrée.

Modification prosodique

La source d'excitation (ou erreur résiduelle) obtenue doit subir les modifications prosodiques nécessaires (durée et pitch) en utilisant la méthode TD-PSOLA expliquée ci-dessus. Les facteurs de modification de la durée et de la fréquence fondamentales sont calculés à chaque instant d'analyse t_a (marque de pitch).

Synthèse

Le signal de parole synthétique est reproduit en faisant passer le signal résiduel à travers le filtre de synthèse.

3.2.4. Modèle harmoniques plus bruit (HNM)

Le modèle HNM (Harmonic plus Noise Model) est basé sur une représentation harmonique plus bruit du signal parole en synchronisme avec le pitch. L'implémentation du modèle HNM passe par une étape d'analyse permettant l'extraction des différents paramètres de la partie harmonique et ceux de la partie bruit du signal parole, une étape de modification des paramètres prosodiques, et enfin une étape de synthèse qui consiste à générer une parole synthétique avec la prosodie désirée.

3.2.4.1 Analyse

Théoriquement, les paramètres HNM peuvent être estimés par une technique d'analyse par synthèse c'est-à-dire, par l'optimisation d'une fonction de coût entre le signal original et le signal synthétique. Cependant, cette approche revient à résoudre analytiquement un problème d'optimisation non linéaire de grande dimension. Pour simplifier le problème d'estimation des paramètres HNM, les paramètres de la partie harmonique et de la partie bruitée sont estimés séparément. L'estimation de la fréquence fondamentale et de la fréquence maximale de voisement est isolée de l'estimation des amplitudes et des phases des harmoniques. Ainsi, la première étape d'analyse consiste à estimer la fréquence fondamentale et la fréquence maximale de voisement pour les trames voisées.

Estimation de la fréquence fondamentale et de la fréquence maximale de voisement

De la bonne détermination de la fréquence fondamentale dépendent aussi les mesures des amplitudes, Il est donc impératif de connaître F_0 d'une façon fiable. De nombreux travaux cherchent à améliorer la robustesse de l'évaluation, dans des contextes de plus en plus drastiques (environnement d'hélicoptères ou dans une foule par exemple) [125-128].

Deux méthodes qui permettent une bonne estimation de la fréquence fondamentale sont décrites dans la référence [126] : la première repose su un critère lié à l'autocorrélation normalisée du signal temporelle (dite méthode MBE) ; la seconde utilise une représentation fréquentielle à court-terme (dite méthode DOVAL). L'estimation précise de la fréquence maximale de voisement F_C (séparant les deux composantes harmonique et bruit d'un signal vocal dans le domaine fréquentiel) est également nécessaire dans un modèle HNM. Dans ce travail, nous utilisons un algorithme qui permet l'estimation du pitch, de la décision du voisement, et de la fréquence maximale de voisement en même temps. Cet algorithme a été adapté par STYLIANOU au modèle HNM dans [129]. L'analyse commence par déterminer un pitch initial $\hat{F}_0$ par une méthode temporelle basée sur une méthode de maximisation de la fonction d'autocorrélation. Ce pitch initial est utilisé par la suite, pour la décision de voisement, pour l'estimation de la fréquence maximale de voisement et finalement pour le raffinement de l'estimation du pitch.

En utilisant les fréquences fondamentales estimées, nous sommes maintenant capable de déterminer les instants d'analyse t_a^i. Dans le cas des fenêtres voisées, ces instants sont synchronisés avec T_0^i, la période fondamentale locale :

$$t_a^{i+1} = t_a^i + T_0^i \tag{3.1}$$

Les fenêtres d'analyse utilisées sont de longueur égale à $2T_0^i$ et sont centrées sur les instants d'analyse t_a^i. Notons que la longueur de ces fenêtres varie entre 20 ms (50 Hz) et 2 ms (500 Hz) en fonction du pitch local de la fenêtre voisée, et que leurs centres sont indépendants des instants de fermeture de la glotte. Notons également que dans le cas des fenêtres non voisées, Les instants d'analyse sont incrémentés par une durée constante de 10 ms :

$$t_a^{i+1} = t_a^i + 10ms \tag{3.2}$$

Estimation des amplitudes et des phases

Pour l'estimation des amplitudes et des phases des harmoniques, deux méthodes existent : la méthode de détection de pics et la méthode des moindres carrés [126]. Dans ce travail, nous avons choisi la dernière méthode.

Afin d'obtenir les valeurs des amplitudes et des phases, la façon la plus intuitive de procéder consiste à minimiser l'erreur quadratique entre le signal original et le modèle choisi. Le modèle harmonique décrit par l'équation (1.9) se réécrit sous la forme :

$$h(n) = \sum_{l=1}^{L} a_l \cos(2\pi l F_0 n) + b_l \sin(2\pi l F_0 n) \tag{3.3}$$

pour $n = n_1 ... n_N$ et l'erreur à minimiser suivant les valeurs de a_l et b_l est :

$$E = \sum_n \left| s(n) - h(n) \right|^2 \tag{3.4}$$

Il s'agit d'un problème de type "moindres carrés" qui se résout donc explicitement par :

$$\underline{\theta} = (H^t H)^{-1} H^t \underline{X} \tag{3.5}$$

avec $\underline{\theta} = (a_1, \ldots, a_L, b_1, \ldots, b_L)^t = (\theta_1, \ldots, \theta_{2L})^t$, $\underline{X} = (s_1, \ldots, s_N)^t$ les échantillons du signal original et H la matrice des $\cos(2\pi l F_0 n)$ et $\sin(2\pi l F_0 n)$.

$$H = \begin{bmatrix} \cos(2\pi F_0 n_1) & \cdots & \cos(2\pi L F_0 n_1) & \sin(2\pi F_0 n_1) & \cdots & \sin(2\pi L F_0 n_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cos(2\pi F_0 n_N) & \cdots & \cos(2\pi L F_0 n_N) & \sin(2\pi F_0 n_N) & \cdots & \sin(2\pi L F_0 n_N) \end{bmatrix} \tag{3.6}$$

L'estimation du vecteur de paramètres $\underline{\theta}$ donne ensuite accès aux amplitudes et aux phases des composantes harmoniques par les relations :

$$A_l = \sqrt{\theta_l^2 + \theta_{l+L}^2} \tag{3.7}$$

et

$$\Phi_l = -\arctan\left(\frac{\theta_{l+L}}{\theta_l} \right) \tag{3.8}$$

Estimation des paramètres du bruit

Pour toutes les trames d'analyse, qu'elles soient voisées ou non, la densité spectrale de puissance est modélisée par un filtre AR tout-pôle d'ordre 10 en utilisant la méthode d'autocorrélation standard [130]. La fonction d'autocorrélation est estimée en utilisant 40 ms du signal centré autour de chaque instant d'analyse. Le gain du filtre est donné par la variance du signal sur la même durée.

3.2.4.2. Estimation des enveloppes de phase et d'amplitude

L'estimation des enveloppes de phase et d'amplitude peut être considérée comme une étape intermédiaire entre l'analyse et la synthèse. Dans la synthèse de la parole, ces enveloppes sont utilisées, par exemple, dans le cas de modifications spectrales et/ou de la fréquence fondamentale [26, 52]. Les amplitudes et les phases calculées lors de l'analyse ne correspondent pas aux amplitudes et aux phases des nouvelles composantes harmoniques. Elles prennent comme valeurs celles que prennent l'enveloppe spectrale et l'enveloppe de phase aux nouvelles fréquences.

L'enveloppe de phase est obtenue par un algorithme de déroulement de phase décrit par STYLIANOU dans [52]. Cette technique permet de préserver la continuité de la phase aussi bien dans le domaine fréquentiel que dans le domaine temporel.

L'enveloppe spectrale correspond au spectre d'amplitude du filtre modélisant le conduit vocal et la partie lisse du spectre de la source glottique. L'enveloppe spectrale est donc une courbe qui passe par les pics des harmoniques sur la partie qui s'étend jusqu'à la fréquence maximale de voisement et qui suit le spectre de la partie bruitée pour les fréquences supérieures à la fréquence maximale de voisement [26].

Estimation de l'enveloppe spectrale

La méthode la plus utilisée pour l'estimation de l'enveloppe spectrale utilise le cepstre discret qui a été introduit par GALAS et RODET[131]. Cette méthode consiste à déterminer les coefficients cepstraux conduisant à une enveloppe spectrale passant le plus proche possible des amplitudes des harmoniques. Etant données les amplitudes des harmoniques A_l d'une trame voisée, les coefficients du cepstre discret $c = \left[c_0 \cdots c_p \right]$, où p est l'ordre du cepstre, sont obtenus en minimisant un critère des moindres carrés :

$$\varepsilon_r = \sum_{l=1}^{L} \left| \log A_l - \log \left| S(f_l, c) \right| \right|^2 , \tag{3.9}$$

où $\left| S(f_l, c) \right|$ est l'amplitude du spectre qui est reliée aux coefficients du cepstre par :

$$\log \left| S(f, c) \right| = c_0 + 2 \sum_{i=1}^{p} c_i \cos(2\pi f i) \tag{3.10}$$

et

$$f_l = l F_0 \tag{3.11}$$

Une fois les c_i calculés, on en déduit $|S(f,c)|$ pour toute fréquence *f*. La solution des moindres carrés est donnée par :

$$c = \left(M^t M\right)^{-1} M^t a \tag{3.12}$$

avec $a = \left[\log(A_1) \cdots \log(A_L)\right]^t$ et

$$M = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 2) & \cdots & 2\cos(2\pi f_1 p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2\cos(2\pi f_L) & 2\cos(2\pi f_L 2) & \cdots & 2\cos(2\pi f_L p) \end{bmatrix} \tag{3.13}$$

Lorsque *L* et *p* sont du même ordre de grandeur, la matrice $M^t M$ est en général mal conditionnée (et elle est singulière si *p*>*L*). Afin de passer outre ce problème de conditionnement, cette méthode a été améliorée par CAPPE [132, 133] en introduisant un terme de régularisation dans l'expression à minimiser :

$$\varepsilon_r = \sum_{l=1}^{L} \left|\log A_l - \log\left|S(f_l,c)\right|\right|^2 + \lambda \Re\left[S(f_l,c)\right] \tag{3.14}$$

où λ est le paramètre de régularisation et la fonction de pénalisation $\Re$ est définie par :

$$\Re\left[S(f,c)\right] = c^t R c \tag{3.15}$$

R est une matrice diagonale donnée par :

$$R = \begin{bmatrix} 0 & & & & \\ & 8\pi^2 1^2 & & & \\ & & 8\pi^2 2^2 & & \\ & & & \ddots & \\ & & & & 8\pi^2 p^2 \end{bmatrix} \tag{3.16}$$

La solution dans ce cas est :

$$c = \left(M^t M + \lambda R\right)^{-1} M^t a \tag{3.17}$$

Une autre amélioration consiste à utiliser une transformation de l'échelle des fréquences, par exemple en échelle de BARK et de MEL, afin de reproduire le plus fidèlement possible les propriétés d'audition de l'oreille [52].

Ainsi, une fenêtre voisée HNM est complètement décrite par sa fréquence fondamentale, le nombre d'harmoniques, les coefficients du cepstre discret, l'enveloppe de

phase, les coefficients du filtre AR, et le gain du filtre (gain LP). Une fenêtre non voisée est représentée uniquement par les coefficients du filtre AR et son gain.

3.2.4.3. Modification et synthèse

La synthèse est une opération consistant à calculer les échantillons du signal de parole à partir des paramètres HNM. Elle est effectuée de façon pitch-synchrone comme l'étape d'analyse. Dans le cas où aucune modification n'est faite sur le signal parole les instants de synthèse, t_s^i , (ou marques de synthèse) coïncident avec les instants d'analyse t_a^i (où marques d'analyse). Par contre si le signal vocal doit subir des modifications prosodiques, il faudra alors calculer les nouveaux instants de synthèse et les nouveaux paramètres HNM.

Modifications prosodiques

Deux tâches principales sont à effectuer lors des modifications prosodiques. La première consiste à estimer les instants de synthèse. La seconde permet la ré-estimation des amplitudes et des phases des harmoniques modifiés.

Etant donné les instants d'analyse, t_a^i , les facteurs de modification du pitch, $\alpha(t)$, et les facteurs de modification de la durée, $\beta(t)$, un algorithme récursif permet de déterminer les instants de synthèse t_s^i [52, 126]. Supposons que le contour de pitch original, *P(t)*, est continu et que l'instant de synthèse t_s^i est connu, les instants de synthèse t_s^{i+1} sont alors donnés par :

$$t_s^{i+1} = t_s^i + \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} \frac{P(t)}{\alpha(t)} dt \tag{3.18}$$

où $t_v^{(\cdot)}$ est un temps virtuel défini par :

$$t_s^i = D(t_v^i) \tag{3.19}$$

où la fonction D(t) est donné par :

$$D(t) = \int_0^t \beta(\tau) d\tau \tag{3.20}$$

Cette fonction nous permet de faire coïncider les instants d'analyse et de synthèse. Les instants virtuels sont définis sur l'axe temporel d'analyse et ne coïncident pas, en général,

avec les instants d'analyse réels. Par conséquent, étant donné un instant virtuel, t_v^i, avec $t_a^i \le t_v^i \le t_a^{i+1}$, il y a deux options : ou bien nous devons interpoler les paramètres HNM à partir de t_a^i et t_a^{i+1}, ou décaler t_v^i vers l'instant d'analyse le plus proche (t_a^i ou t_a^{i+1}). Dans cette implémentation, nous avons utilisé la seconde option. Les intégrales dans les équations (3.18) et (3.20) peuvent être facilement approximés si *P(t)*, $\alpha(t)$, et $\beta(t)$, sont supposés constants par morceaux.

Une fois les instants de synthèse sont déterminés, l'étape suivante est l'estimation des amplitudes et des phases aux nouvelles fréquences harmoniques. Ceci est fait en échantillonnant l'enveloppe spectrale et l'enveloppe de phases aux nouvelles fréquences.

Synthèse

La synthèse se fait de façon pitch-synchrone en utilisant le processus d'addition-recouvrement (OverLap and Add). Pour la synthèse de la partie harmonique d'une fenêtre, l'équation (1.9) est utilisée. La partie bruit est obtenue en filtrant un bruit blanc gaussien par un filtre AR tout-pôle. Si la fenêtre est voisée, la partie bruit est filtrée par un filtre passe haut de fréquence de coupure égale à la fréquence maximale de voisement. Ensuite, elle sera modulée par une enveloppe temporelle synchronisée avec la période de pitch (voir figure 3.1) [52].

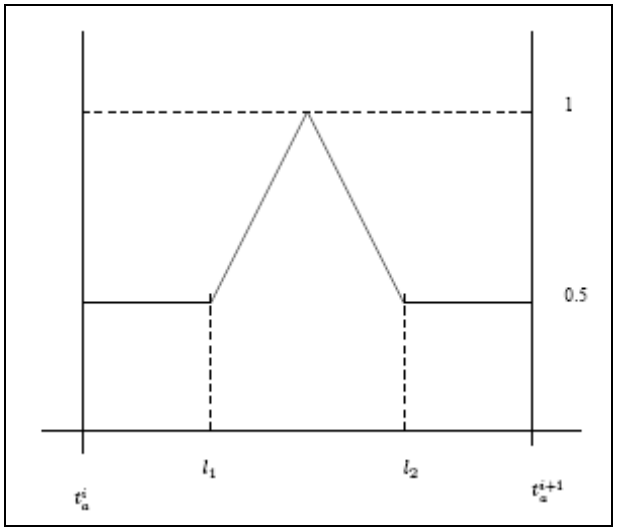


Figure 3.1 : Fonction de modulation temporelle du bruit. t_s^i et t_s^{i+1} sont deux instants de synthèse successifs. $l_1 = 0.15(t_s^{i+1} - t_s^i)$ et $l_2 = 0.85(t_s^{i+1} - t_s^i)$

3.3. Marquage de F_0 et estimation d'un chemin optimal

Rappelons que l'implémentation des deux techniques TD-PSOLA et LP-PSOLA nécessite une estimation précise des instants de fermeture de la glotte (marques de pitch). Dans notre cas, nous avons utilisé une méthode de marquage de pitch basée sur la programmation dynamique [120]. Le principe de la programmation dynamique, ainsi que l'algorithme de placement des marques de pitch seront décrits dans cette section.

3.3.1. Principe de la programmation dynamique

La programmation dynamique doit être utilisée pour calculer un chemin optimal, de la première à la dernière colonne d'une matrice rectangulaire avec une contrainte sur la pente maximale admissible S_{max} [120]. Etant donnée une matrice A de dimensions (N, M), nous définissons un vecteur de chemin optimal p dont les composantes correspondent à l'indice d'une ligne pour chaque colonne de A. Soit E la somme des éléments de A le long du chemin optimal p.

$$E = \sum_{m=1}^{M} A(p(m), m) \tag{3.21}$$

L'algorithme de programmation dynamique consiste à trouver le vecteur p pour lequel E est maximale sous la contrainte de pente maximale admissible S_{max} pour le chemin optimal.

$$\left| p(m) - p(m-1) \right| \leq S_{\max} \tag{3.22}$$

3.3.2. Placement des marques de pitch

Soit pitch(n) le vecteur dont les composantes sont les périodes de pitch estimées par l'algorithme d'extraction de F_0, pour chaque échantillon du signal vocal d'entrée S(n) (Npts points totaux). L'algorithme de marquage passe par les étapes suivantes :

- Nous devons initialiser les marques de pitch (en se basant sur un placement arbitraire de la première marque à n=1).

```
marque_mat ←—— Npts zeros;   n=1
tant que (n ≤ Npts)
marque_mat(n)=1
n=n+pitch(n)
fin
```

- Nous désirons localiser ces marques de pitch aux instants de fermeture de la glotte (pics maximum du signal vocal) pour ne pas avoir de déphasage entre le signal synthétique et le signal original. Pour cela, nous considérons le vecteur C(k) renfermant les valeurs de n pour lesquelles marque_mat(n)=1.

- Nous définissons la matrice B comme suit : chaque colonne k de la matrice B a pour éléments les valeurs

$$\left|S\left(C(k)+\left[-P_{\max}..P_{\max}\right]\right)\right| * \text{Fenêtre de Hanning} \qquad (3.23)$$

P_{max} étant la valeur maximale admissible de la période de pitch (exprimé en nombre d'échantillons). Elle est donnée par :

$$P_{\max} = round(f_s/60) \quad \text{Échantillons} \qquad (3.24)$$

f_S est la fréquence d'échantillonnage du signal.

- L'étape finale consiste à déterminer, par programmation dynamique comme c'est expliqué plus haut, le vecteur du chemin optimal p de la matrice B. Ainsi, les marques de pitch désirées sont :

$$\text{marques\_pitch}(k)=C(k)+p(k)$$

k représente la $k^{ème}$ marque de pitch.

3.4. Résultats d'implémentation

La qualité de la synthèse par TD-PSOLA et LP-PSOLA dépend fortement de la bonne localisation des marques de pitch du signal vocal. Pour mettre en évidence l'efficacité de l'algorithme de marquage utilisé, nous avons présenté sur la figure 3.2 une tranche voisée d'un signal vocal au dessous de laquelle nous avons présenté des barres verticales indiquant les marques de pitch obtenus.

L'implémentation des quatre modèles a été prévue pour être employée afin de concaténer des segments de parole extraits d'une base de données. Cependant, en traitant une expression entière comme un seul segment, renfermant un certain nombre de transitions de voisement, il est possible de démontrer que l'algorithme fonctionne correctement. Ceci est également utile dans la détermination du degré de dégradation introduit par les modifications prosodiques seules.

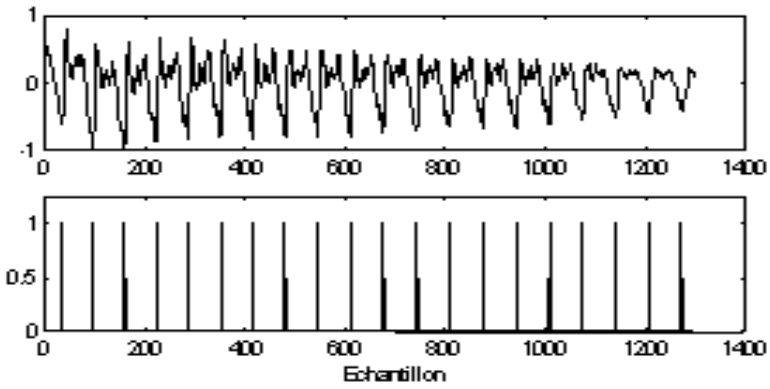


Figure 3.2 : Marquage des périodes. De haut en bas: une tranche voisée d'un signal vocal, et la position des marques d'analyse

La figure 3.3 présente des modifications de la fréquence fondamentale (intonation) et de la durée par la technique TD-PSOLA. Nous remarquons que les différentes modifications sont bien réussies. En écoutant la parole synthétique produite, nous constatons que cette parole est de bonne qualité dans le cas d'augmentation de F_0 d'un facteur inférieur ou égal à 2, ou d'une diminution d'un facteur de 1.5 ou moins. Dans le cas de modification de la durée, la parole synthétique est de haute qualité avec une légère dégradation audible dans les segments rallongés. Cette dégradation est due à la périodicité artificielle introduite dans les régions non voisées de la parole synthétique. Cependant, cette dégradation est prévue, et de telles grandes augmentations dans la durée n'étaient pas susceptibles d'être exigées par le système de synthèse. En effet, le facteur d'augmentation de la durée ne doit pas dépasser 1.5, et celui de réduction de la durée doit rester inférieur ou égal à 2.

Sur les figures de 3.4 à 3.8 nous proposons d'étendre les deux techniques TD-PSOLA et LPC pour un facteur de modification de l'intonation variant dans le temps. Pour cela, nous avons utilisé l'outil graphique pour afficher le contour mélodique du signal original et donner à l'utilisateur la possibilité de tracer l'allure de la courbe mélodique désirée en utilisant la souris.

La figure 3.4 illustre le contour mélodique du mot "parenthèse" (prononcé par un locuteur masculin sous une forme déclarative), et l'intonation (représentée par des étoiles) que nous désirons obtenir au niveau du signal synthétique après synthèse par la technique TD-PSOLA. Nous avons choisi la courbe mélodique qui commence par croître vers la fin du signal, afin d'avoir la prononciation sous une forme interrogative (nous pouvons obtenir

n'importe quelle forme de la phrase : affirmative, exclamative, interrogative avec les différentes élocutions, etc.).

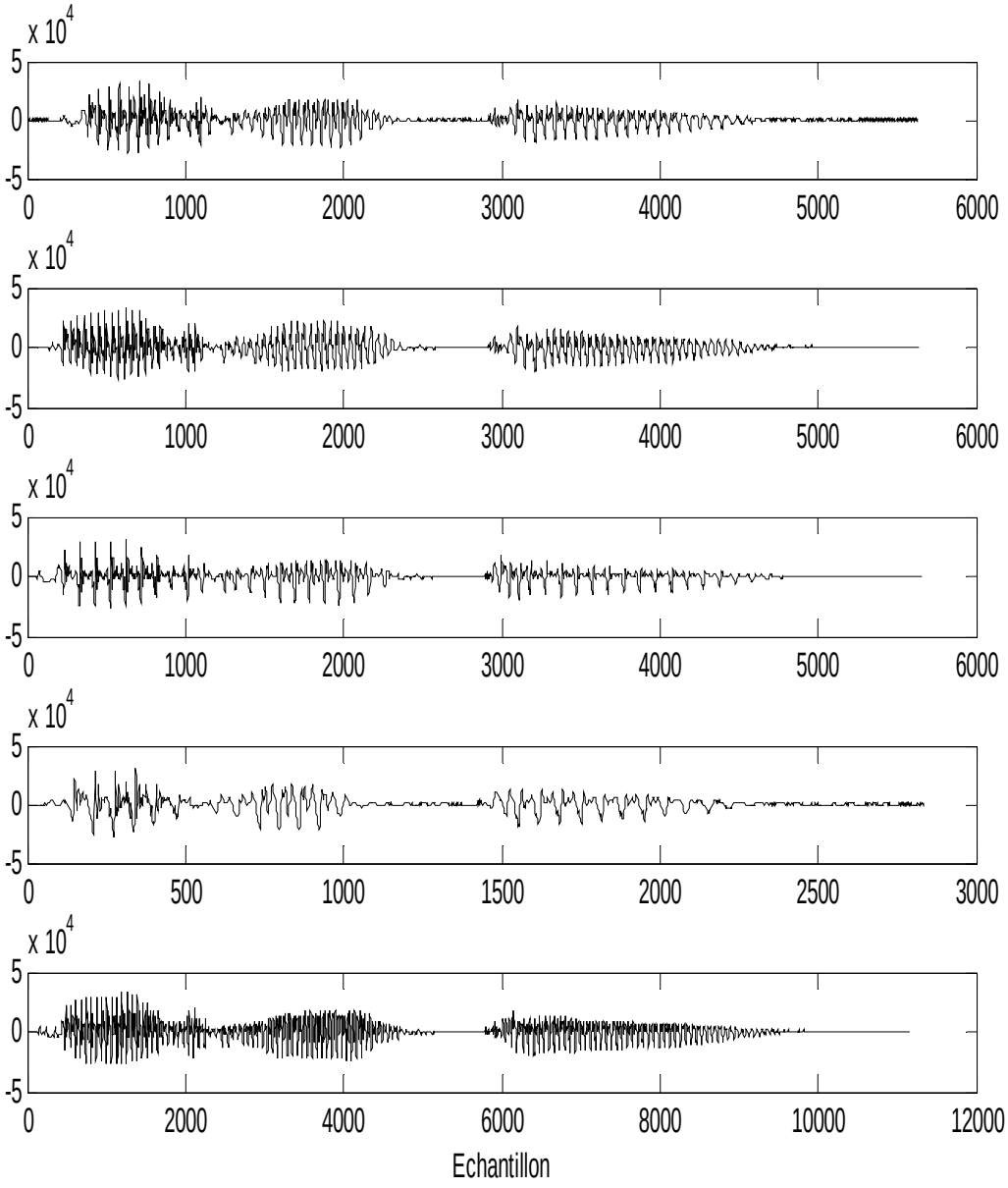


Figure 3.3 : Analyse-synthèse par TD-PSOLA. De haut en bas nous avons : le signal original ; le signal synthétique avec une augmentation de F_0 d'un facteur 1.5 ; diminution de F_0 d'un facteur de 1.5 ; diminution de la durée d'un facteur 2 ; augmentation de la durée d'un facteur 2.

Sur la figure 3.5 nous proposons le signal original, la fréquence fondamentale de ce signal et la fréquence désirée calculée par interpolation pour chaque échantillon. Sur la figure 3.6 nous avons présenté le contour mélodique désiré (représenté par des étoiles) et celui obtenu par l'algorithme d'extraction de la fréquence fondamentale après modification par la technique TD-PSOLA.

Grâce à cette figure nous avons pu mettre en évidence l'efficacité de la technique TD-PSOLA pour la modification des paramètres prosodiques et la synthèse du signal vocal. Nous voyons bien que la fréquence fondamentale du signal synthétique suit parfaitement la fréquence désirée. Sur le plan perceptuel, en écoutant le fichier synthétique, nous entendons le mot "Parenthèse" prononcé sous une forme interrogative : Parenthèse?

Une autre consigne de la fréquence fondamentale F_0 à suivre, est donnée au niveau de la figure 3.7 Nous remarquons que la fréquence F_0 du signal synthétique suit bien la fréquence désirée. La figure 3.8 montre le signal original et le signal synthétique après cette modification de l'intonation par la méthode TD-PSOLA.

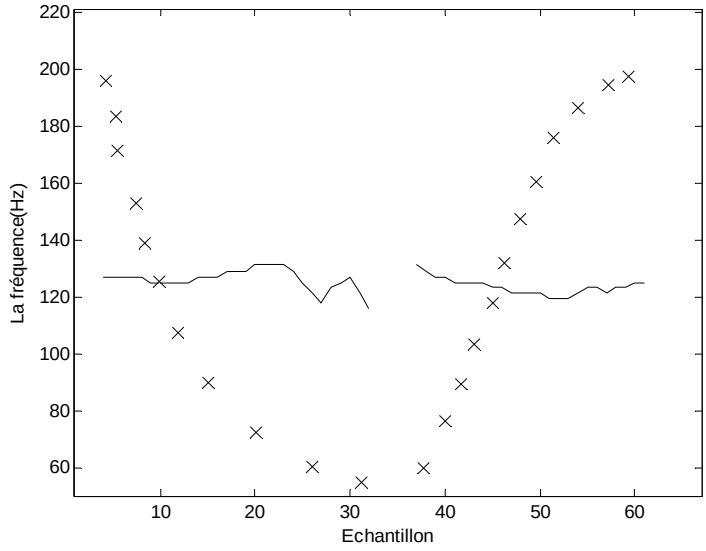


Figure 3.4 : Le contour mélodique du signal original et le contour désiré tracé sous forme d'étoiles par des clicks de la souris

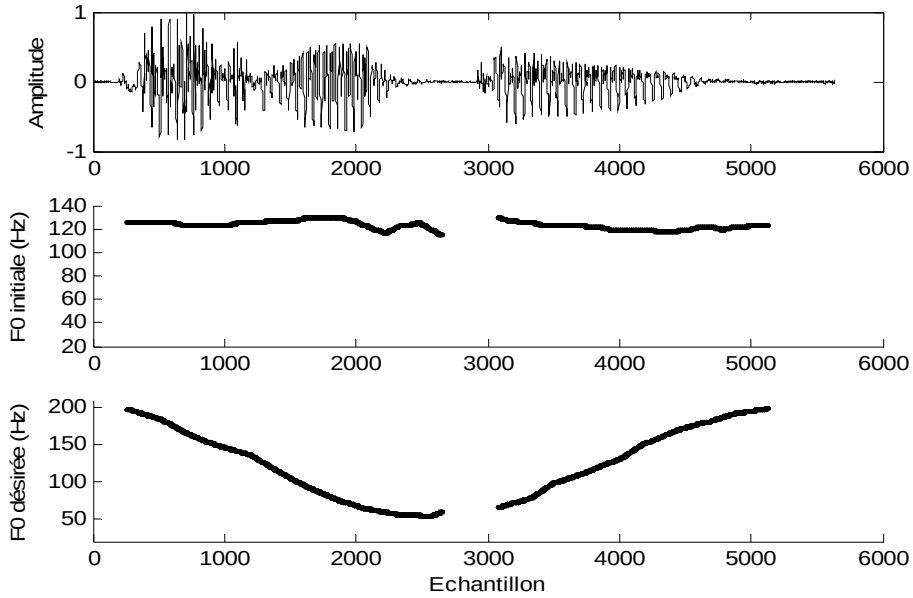


Figure 3.5 : Signal original, le contour mélodique original et le contour mélodique désiré après interpolation sur tous les échantillons

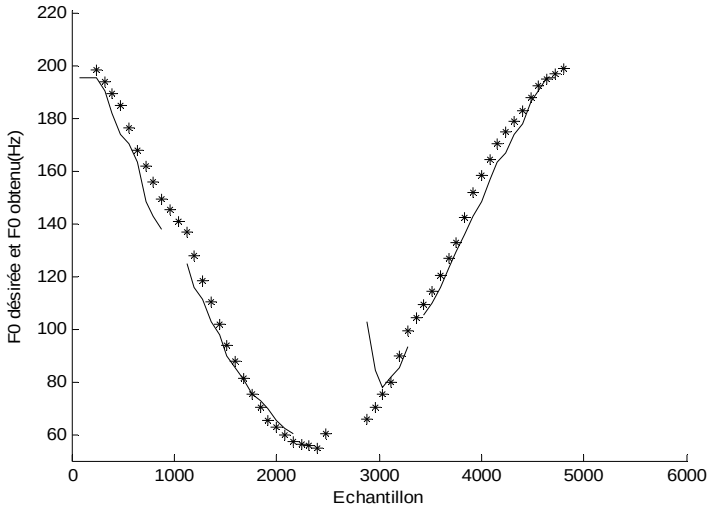


Figure 3.6 : La fréquence fondamentale désirée (représentée sous forme d'étoiles) et la fréquence fondamentale obtenue après modification par TD-PSOLA (en continu)

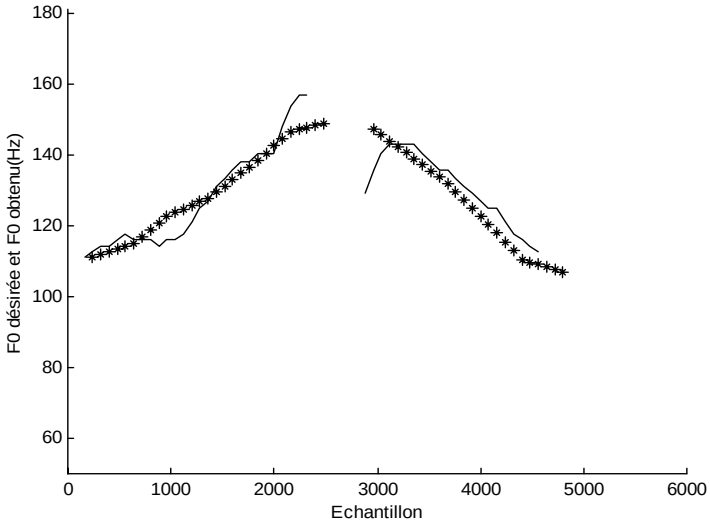


Figure 3.7 : F_0 désirée (représentée sous forme d'étoiles) et F_0 obtenue après modification par la méthode TD-PSOLA (en continu)

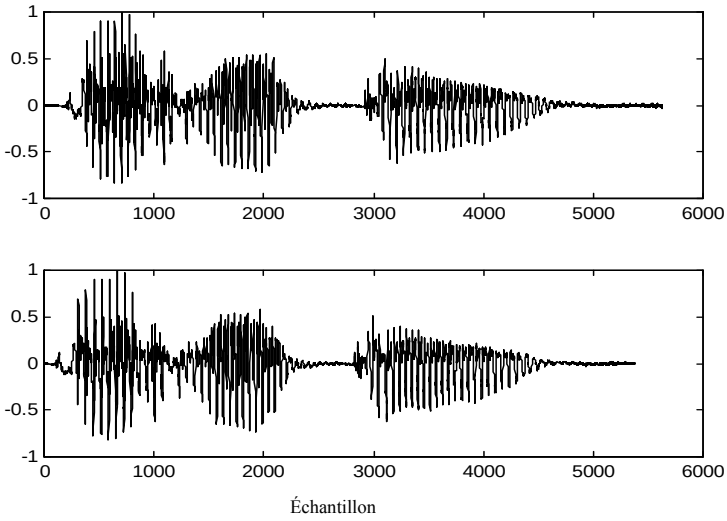


Figure 3.8 : Le signal original et le signal synthétique après modification de la fréquence fondamentale (de la figure 3.7) par la méthode TD-PSOLA

Quelques résultats d'analyse, modification et synthèse par la technique LPC sont représentés dans les figures 3.9 et 3.10. Nous remarquons que la technique LPC offre une très bonne poursuite du contour mélodique désiré. Ceci est dû au fait que F_0 est un paramètre explicite du modèle LPC. La parole synthétique produite est loin d'être parfaite, particulièrement pour les parties voisées à cause de l'excitation simplifiée (sous forme de trains d'impulsions). Donc nous avons écarté l'utilisation du modèle LPC dans le système TTS que nous désirons construire pour l'Arabe.

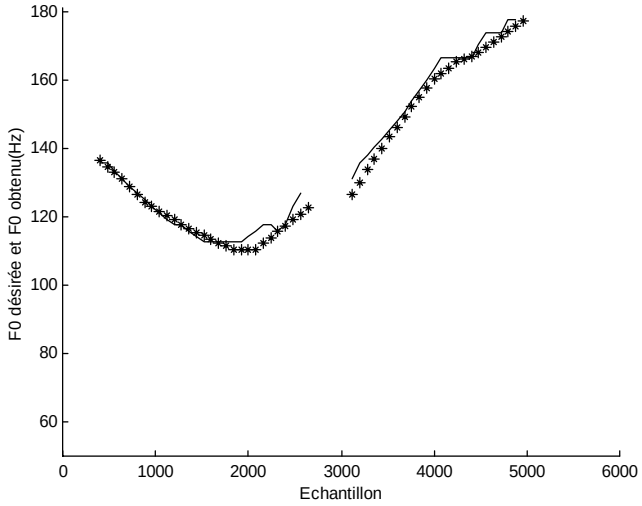


Figure 3.9 : la F_0 désirée (en étoiles) et la F_0 obtenue après modification par la méthode LPC (en continu)

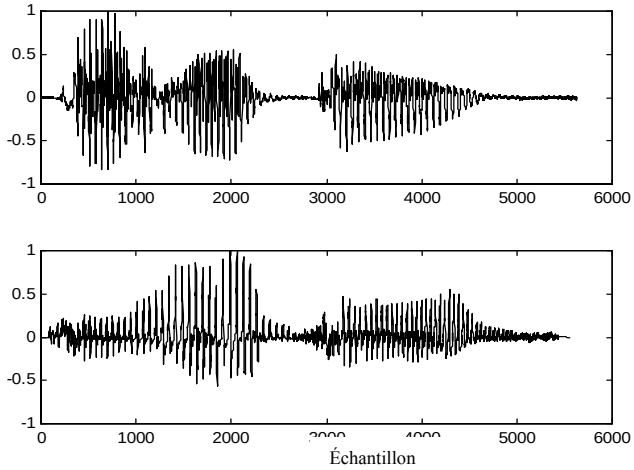


Figure 3.10 : Le signal original et le signal synthétique après modification de F_0 par la méthode LPC

Pour tester les performances de l'algorithme TD-PSOLA dans un système TTS, nous l'avons utilisé comme étage de sortie dans un système de synthèse que nous avons réalisé. Ce système est basé sur une simple concaténation de phonèmes en tenant compte de leurs contextes phonétiques gauche et droit (ce système sera décrit dans la Section 4.6).

Nous avons intégré la méthode TD-PSOLA dans ce synthétiseur pour imposer l'intonation demandée de la phrase à synthétiser, et allonger ou réduire chaque phonème à la durée désirée. Il est à noter que dans ce cas, le facteur de modification de l'intonation n'est pas fixe, il varie avec le temps. Le facteur de modification de la durée quant à lui , est

constant durant toute la durée d'un phonème, mais varie d'un phonème à un autre selon la structure rythmique désirée.

La figure 3.11 montre l'onde temporelle et l'intonation de la phrase arabe naturelle :

"وَ تُوُفِّيَ عَامَ ألفٍ و َ ثَمَان مِئَةٍ وَ ثَلاثَةٍ وَ ثَمَانِين" enregistrée par un locuteur Jordanien avec une fréquence d'échantillonnage f_s=16kHz et une résolution de 16 bits. Nous avons fait l'alignement de cette phrase pour déterminer la durée de chaque phonème. Ensuite, nous avons généré cette même phrase par notre synthétiseur en lui imposant l'intonation et les durées phonétiques de la phrase naturelle. Le résultat obtenu est représenté sur la figure 3.12. Nous remarquons que l'allure générale de l'intonation suit bien celle de la phrase naturelle.

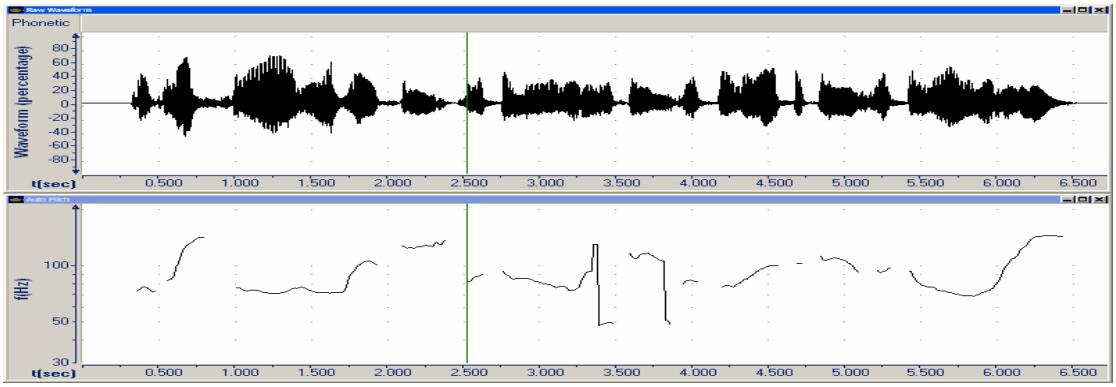


Figure 3.11 : L'onde temporelle et l'évolution de la fréquence fondamentale de la phrase naturelle arabe :" وَ تُوُفِّيَ عَامَ ألفٍ و َ ثَمَان مِئَةٍ وَ ثَلاثَةٍ وَ ثَمَانِين "

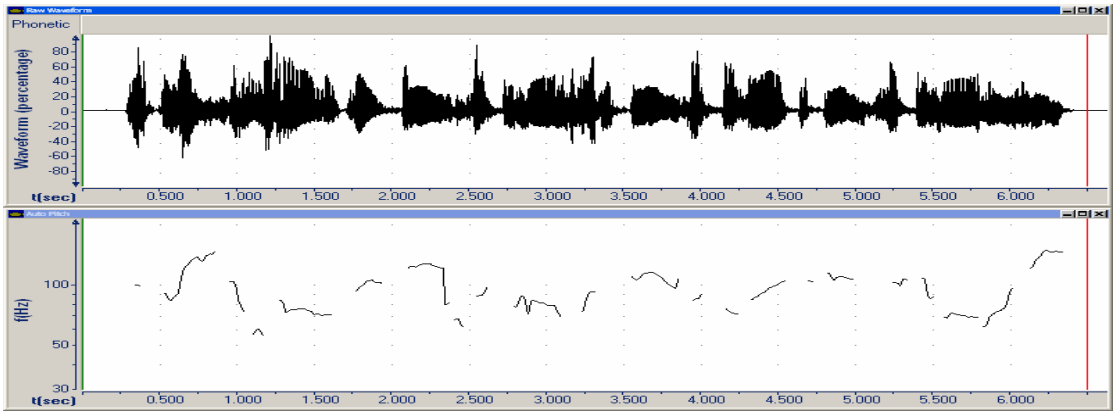


Figure 3.12 : L'onde temporelle et l'évolution de F_0 de la phrase synthétique arabe :

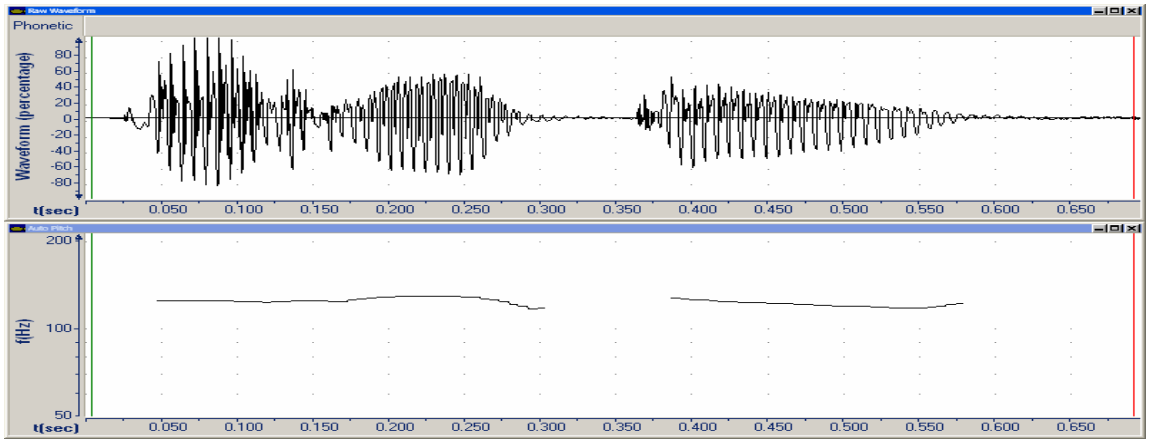
" وَ تُوُفِّيَ عَامَ ألفٍ و َ ثَمَان مِئَةٍ وَ ثَلاثَةٍ وَ ثَمَانِين "

La performance de la méthode LP-PSOLA est testée dans deux contextes :

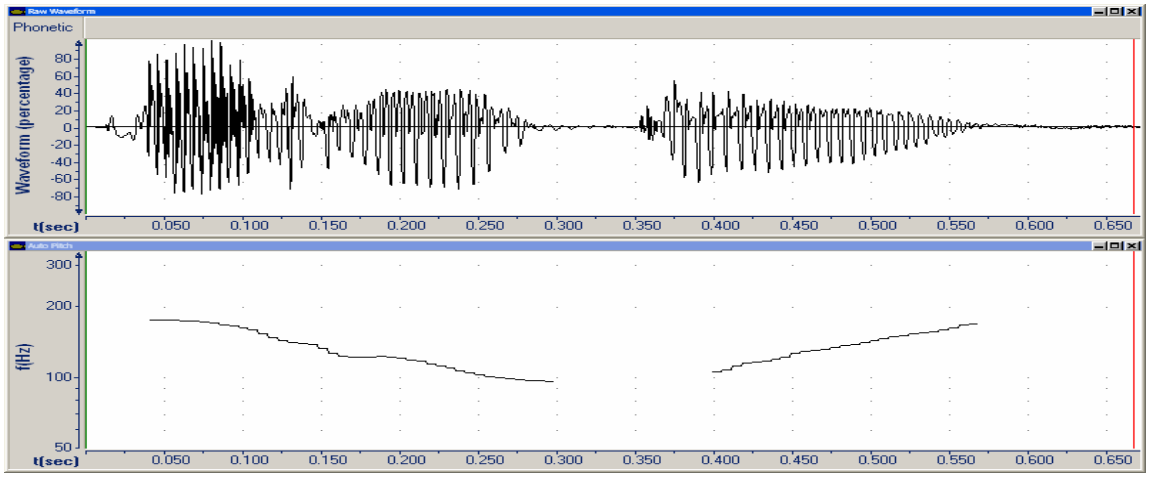
- Dans le premier contexte, nous avons pris deux phrases identiques naturelles prononcées par le même locuteur mais avec des intonations différentes. Nous avons imposé à une phrase la prosodie de l'autre phrase. Le résultat est montré sur la figure 3.13. La figure 3.13 (a) représente le signal naturel source et son intonation. La figure 3.13 (b) illustre le signal naturel cible, alors que sur la figure 3.13(c) nous présentons le signal obtenu après application de la technique LP-PSOLA. Nous remarquons que l'intonation du signal obtenu est très proche du signal cible.

- Dans le deuxième contexte, nous avons intégré la technique LP-PSOLA dans un autre synthétiseur à base de réseaux de neurones, que nous avons réalisé pour l'Arabe Standard (ce synthétiseur est détaillé dans le chapitre 5). Ce synthétiseur convertit les données linguistiques obtenues à la sortie d'un étage de transcription phonétique et traitements linguistique, et la durée segmentale désirée pour chaque phonème, en une série de paramètres d'un codeur LPC qui convertit ces paramètres en parole. LP-PSOLA est utilisé pour modifier les caractéristiques prosodiques de l'excitation qui va être filtrée par le codeur LPC afin d'obtenir une parole synthétique naturelle. La figure 3.14 (a) donne la parole générée à la sortie de notre synthétiseur pour la phrase arabe "الدرس العاشر". Les contours prosodiques imposés à cette phrase sont extraits de la même phrase prononcée par le locuteur Jordanien. En comparant l'intonation et le rythme du signal synthétique obtenu avec ceux de la phrase originale naturelle (figure 3.14 (b)), nous remarquons que les résultats sont satisfaisants.

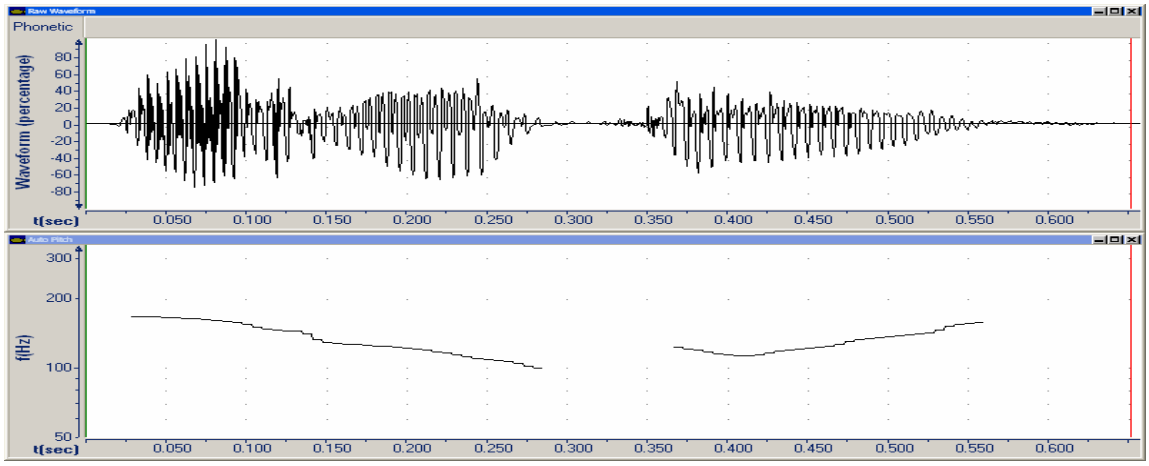
Les résultats de l'implémentation du modèle HNM sont présentés sur les figures (3.15 et 3.16). La figure 3.15 montre le signal parole original d'une phrase, la partie harmonique de cette phrase synthétisée sans modification par le modèle HNM, la partie bruit, et le signal résultant de l'addition des deux parties harmonique et bruit. Il est clair à partir de cette figure que le signal résultant et le signal original sont similaires. Ils sont presque indiscernables perceptuellement. La figure 3.18 montre que ce modèle offre une très bonne adaptation de l'intonation.



(a)

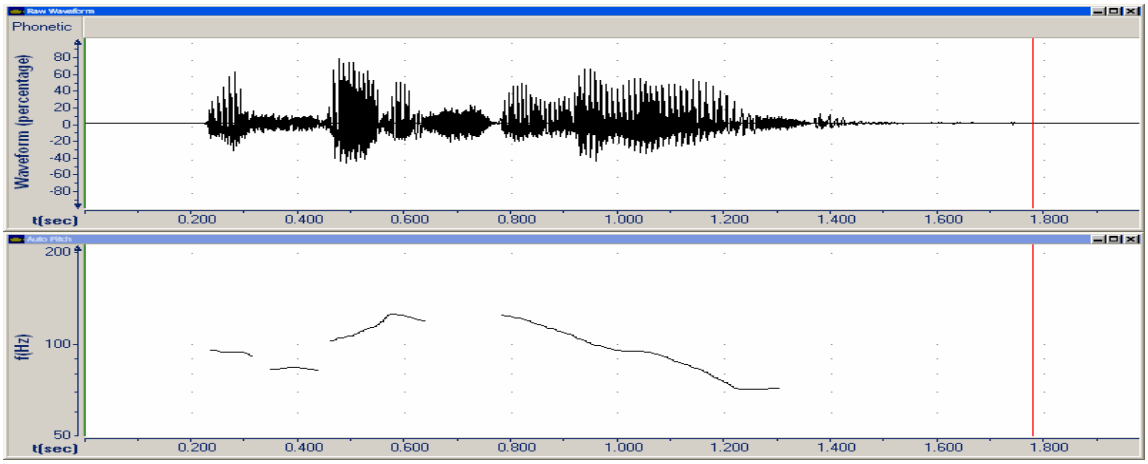


(b)

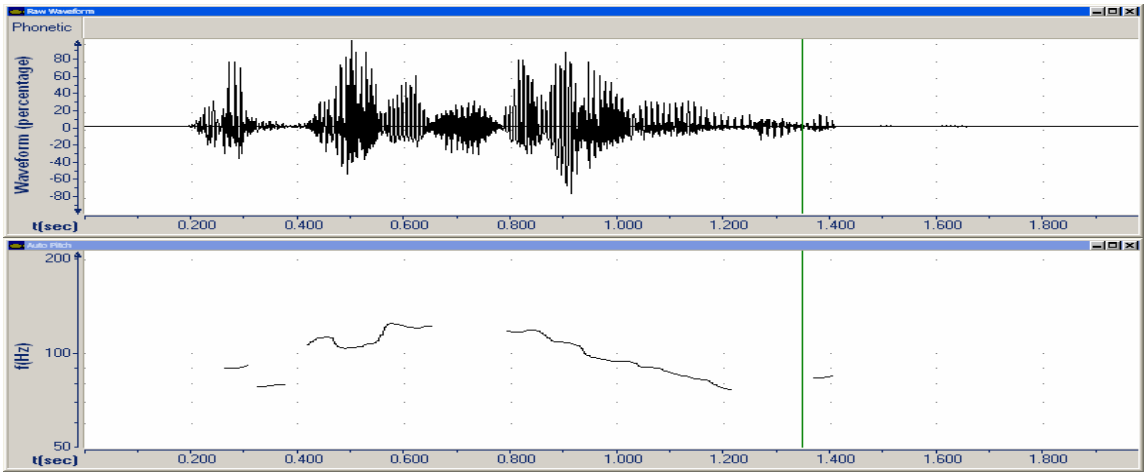


(c)

Figure 3.13 : Modification de la prosodie par LP-PSOLA appliquée à un signal naturel
(a): parole naturelle source, (b): parole naturelle cible (c): parole modifiée



(a)



(b)

Figure 3.14 : Application de LP-PSOLA au synthétiseur vocal à base de réseaux de neurones pour la phrase Arabe : **"الدرس العاشر"** (a) parole naturelle, (b) parole produite par le synthétiseur

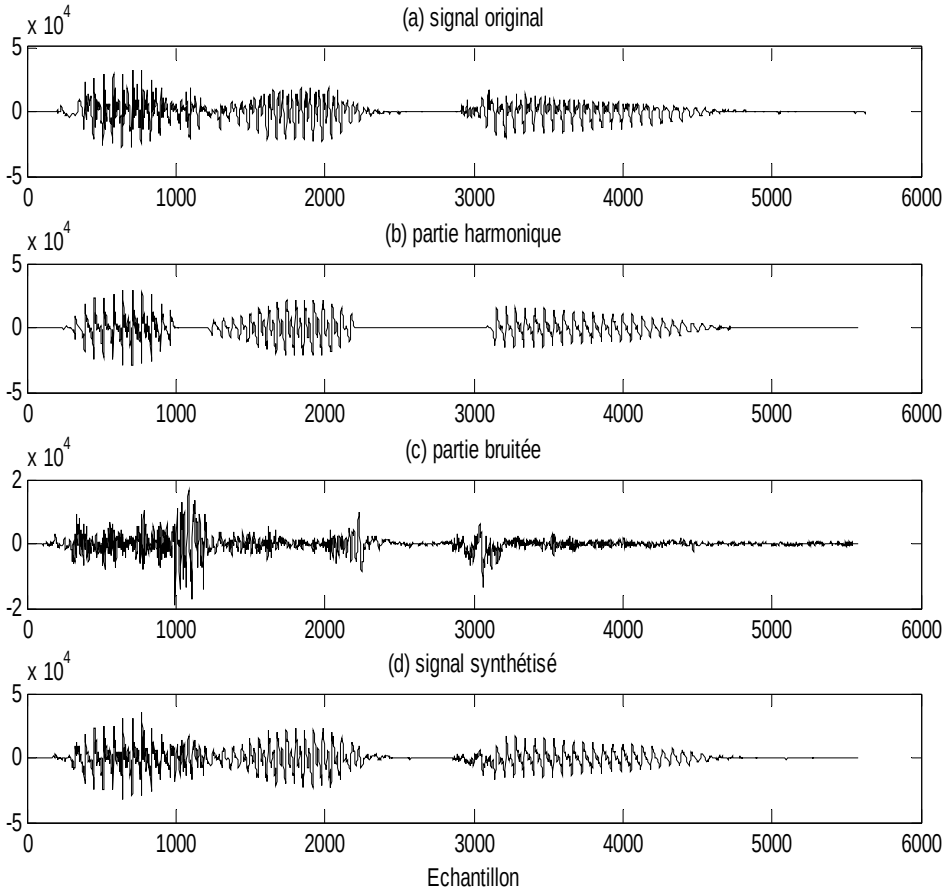


Figure 3.15 : Analyse et synthèse par HNM sans modification de la prosodie

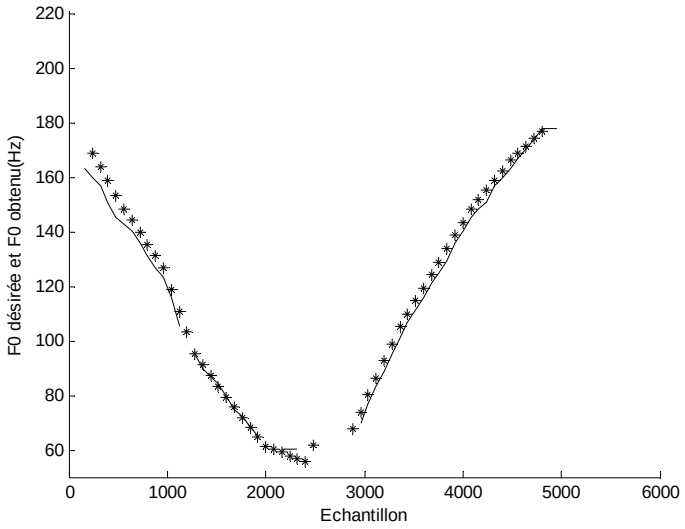


Figure 3.16 : La fréquence fondamentale désirée (représentée sous forme d'étoiles) et la fréquence fondamentale obtenue après modification par le modèle HNM (en continu)

3.5. Conclusion

Dans un effort de choisir une représentation du signal parole pour notre système de synthèse TTS, nous avons fait l'étude des quatre modèles de parole : LPC, TD-PSOLA et LP-PSOLA, et HNM. L'adaptation de la prosodie donne des résultats comparables pour les quatre systèmes. Pour ce qui est des possibilités de concaténation de segments, qui constituent un critère essentiel dans un système TTS, La synthèse LPC offre la possibilité de concaténation par interpolation linéaire de ses coefficients dans le voisinage des points de concaténation. Alors que la TD-PSOLA n'offre que des possibilités limitées de lissage car c'est un modèle non paramétrique. En plus de la disponibilité d'un algorithme de compression efficace et assuré pour le système LPC par rapport à TD-PSOLA. Cependant, le modèle LPC ne peut pas être utilisé dans un système TTS, car il donne une parole synthétique dégradée. Par contre, la TD-PSOLA conduit à une très bonne qualité segmentale car cet algorithme ne fait implicitement usage d'aucun modèle de parole et ne commet par conséquent aucune erreur de modélisation.

Le modèle LP-PSOLA hybride les deux techniques LPC et TD-PSOLA et tire profit de leurs avantages. C'est un modèle paramétrique, ce qui offre la possibilité de lissage spectrale aux points de concaténation des segments. Il est facile à implémenter et permet aussi de générer une parole synthétique de bonne qualité. Le Modèle HNM est lui aussi un modèle paramétrique qui permet d'effectuer des traitements du signal de parole de haute qualité, en particulier des modifications du pitch et de la durée du signal vocal montrent ainsi son utilité dans le cadre de la synthèse de la parole. Sa structure paramétrique lui permet aussi un codage de la parole. Le principal défaut de cette méthode reste néanmoins sa charge de calcul importante liée à sa complexité.

Dans ce travail, nous avons choisi le modèle LP-PSOLA pour réaliser un système TTS à base de réseaux de neurones, sachant qu'un réseau de neurones doit être entraîné pour être en mesure de générer les paramètres de ce modèle (d'où la nécessité d'un modèle paramétrique).

CHAPITRE 4
DEVELOPPEMENT D'UNE BASE DE DONNEES PROSODIQUES POUR L'ARABE STANDARD

4.1. Introduction

La qualité d'un système de synthèse vocale à partir du texte dépend de l'intelligibilité, et du naturel de la parole générée. D'où la nécessité d'un générateur de prosodie automatique de qualité. Parmi les développements les plus récents dans ce domaine, on trouve un intérêt grandissant pour les techniques d'apprentissage automatique, telles que les réseaux de neurones, les arbres de régression et de classification, les modèles de Markov cachés, et d'autres méthodes stochastiques.

L'ensemble de ces techniques se base sur l'analyse de corpus de parole préalablement étiquetés prosodiquement. Vu le manque d'une telle base de données pour la langue Arabe [134, 135], notre objectif est de réaliser une base de sons pour l'Arabe comme celles disponibles pour les autres langues (la base de données TIMIT pour l'Anglais et BDSONS pour le Français, etc.)

4.2. Description du corpus d'analyse utilisé

Le corpus que nous avons utilisé se compose de 148 phrases, comprenant chacune en moyenne 9 mots. Il totalise 1487 mots, 3209 syllabes, 6787 phonèmes dont 2302 voyelles brèves, 561 voyelles longues, 445 semi-voyelles [w] et [y], 1006 consonnes fricatives, 1092 consonnes plosives, 692 consonnes liquides [r] et [l] et 689 consonnes nasales [m] et [n]. Les pauses ont été étiquetées par un expert en apposant le signe « _ » sur le texte correspondant à la voix naturelle [10].

Ces phrases ont été lues à une vitesse moyenne (de 10 à 12 phonèmes/secondes) par un locuteur jordanien, qui n'a reçu aucune consigne particulière afin d'éviter toute influence susceptible d'altérer sa spontanéité. Elles ont été échantillonnées à 16 kHz avec 16 bits par échantillon.

4.3. Analyse du corpus

Nous avons effectué une étude statistique de notre corpus. Le tableau 4.1 montre les résultats de cette étude. Nous pouvons noter les remarques suivantes :

- La fréquence d'occurrence du sukuun est aux alentours de 53.37 % ;
- la voyelle courte [a] et la longue voyelle [aa] apparaissent avec une fréquence de 28%, suivie des voyelles [i] et [ii] avec une fréquence d'occurrence de 14.3%. Les voyelles [u] et [uu] représentent 4.25% ;
- La fréquence d'occurrence des consonnes [?] et [l] (15.94% pour les deux ; 6.11% pour [?] et 9.83% pour [l]) est plus grande que celles des autres consonnes, ceci est dû en partie à leur présence dans les mots comme article de détermination ;
- Les consonnes Arabes les plus fréquentes sont : [l] (9.8329%), [m] (9.54%), [t] (7.87%), [n] (6.92%), [w] (6.53%), [?] (6.11%), et [r] (6.07%), etc.

Tableau 4.1 : Fréquence d'occurrence des voyelles et des consonnes dans le corpus utilisé (%)

Consonnes		**Voyelles**							
		[a]	**[i]**	**[u]**	**[aa]**	**[ii]**	**[uu]**	**sukun**	**Total**
[?]	ءاأ	1.9326	0.8488	0.1567	0.0653	0.0261	0.0522	3.0295	6.1113
[b]	ب	0.8357	0.9271	0.1436	0.1959	0.1306	0.0131	2.2199	4.4659
[t]	ت	1.0838	1.8020	0.3134	0.1828	0.1698	0.0131	4.3092	7.8741
[T]	ث	0.6007	0.1045	0.0261	0.1306	0.0522	0.0000	0.9402	1.8543
[Z]	ج	0.2350	0.0783	0.0131	0.0392	0.0000	0.0000	0.4179	0.7835
[x]	خ	0.1567	0.1175	0.0131	0.0522	0.0522	0.0000	0.3787	0.7704
[X]	ح	0.4962	0.1175	0.1436	0.0653	0.0392	0.0000	0.8488	1.7106
[d]	د	0.8749	0.6921	0.2612	0.2742	0.3134	0.0261	2.6247	5.0666
[D]	ذ	0.2089	0.0522	0.0000	0.1436	0.0000	0.0000	0.3003	0.7051
[r]	ر	1.2405	0.8880	0.2873	0.3003	0.1698	0.0783	3.0948	6.0721
[Z]	ز	0.1306	0.0131	0.0261	0.0653	0.0131	0.0000	0.3656	0.6137
[s]	س	0.5876	0.4048	0.1698	0.1567	0.0914	0.0522	2.5725	4.0350
[S]	ش	0.2350	0.0522	0.0522	0.0522	0.0261	0.0000	0.5746	0.9924
[s.]	ص	0.2089	0.0914	0.0522	0.0522	0.0000	0.0261	0.7182	1.1491
[d.]	ض	0.1175	0.0392	0.0131	0.0261	0.0261	0.0000	0.2220	0.4440
[t.]	ط	0.2873	0.0653	0.0522	0.1175	0.0000	0.0131	0.5615	1.0969
[z.]	ظ	0.0522	0.0783	0.0000	0.0131	0.0392	0.0000	0.1436	0.3265
[H]	ع	1.6845	0.4048	0.2089	0.4440	0.0522	0.0000	2.6639	5.4583
[G]	غ	0.1306	0.0000	0.0131	0.0261	0.0000	0.0000	0.2873	0.4570
[f]	ف	0.4570	1.1622	0.1045	0.0653	0.4440	0.0000	1.9326	4.1656
[q]	ق	0.6137	0.2350	0.0131	0.1698	0.0392	0.0131	1.0708	2.1546
[k]	ك	0.6660	0.0914	0.1045	0.2481	0.0131	0.0522	1.0708	2.2460
[l]	ل	1.8282	0.7704	0.2612	0.7313	0.1306	0.0522	6.0590	9.8329.
[m]	م	2.1416	1.3581	0.7051	0.5354	0.0914	0.0522	4.6618	9.5456
[n]	ن	1.1100	0.7574	0.0914	0.2089	0.2220	0.0261	4.5051	6.9209
[h]	ه	0.8618	0.4832	0.3134	0.6137	0.0783	0.0392	1.9849	4.3876
[w]	و	2.7292	0.1306	0.1045	0.2089	0.0653	0.0000	3.2907	6.5291
[j]	ي	1.1361	0.1959	0.1045	0.2089	0.0522	0.0131	2.5202	4.2309
Total		22.6430	11.9613	3.7477	5.3931	2.3374	0.5223	53.3690	100.00

4.4. Segmentation et étiquetage du corpus

Notre corpus de parole continue a été segmenté et étiqueté par une procédure semi-automatique, qui passe par les étapes suivante [136]:

- une étape de transcription orthographique-phonétique manuelle de chaque phrase en utilisant le système de transcription SAMPA ;
- une segmentation automatique par l'outil MBROLIGN [137, 138] disponible gratuitement sur Internet ;
- une correction manuelle en utilisant l'outil WAVSURFER et la Fonction de Variation Spectrale SVF qui présente des pics dans les zones instables du signal vocal. Ces pics peuvent correspondre aux zones de transitions entre les phonèmes.

4.4.1. Segmentation automatique du corpus par MBROLIGN

L'outil MBOLIGN nécessite comme entrées le fichier en format « wav » de la phrase à segmenter, et un fichier texte renfermant la transcription phonétique de la même phrase. En sortie, il nous donne les courbes, exemple de la figure 4.1 :

- la première courbe représente l'évolution temporelle du signal vocal de la phrase à segmenter ;
- juste au dessous, la figure montre les marques qui indiquent le début et la fin de chaque phonème au sein de la phrase.
- la quatrième courbe illustre l'évolution temporelle de l'intonation et sa version stylisée ;
- alors que la troisième courbe présente le signal vocal de la même phrase synthétisé en lui imposant la version stylisée de l'intonation.

MBROLIGN nous donne aussi un fichier phonétique (.pho) qui renferme les commandes pour le synthétiseur de haute qualité MBROLA (Multi-Band Resynthesis OverLap Add) [139]. Ce fichier est composé de la liste des phonèmes de la phrase avec leurs caractéristiques prosodiques.

Exemple : Une partie du fichier phonétique donné par MBROLIGN pour la phrase arabe :

« بحيث يصعب على العدوّ الوصول إليها » est présentée ci-dessous :

```
_ 210  0   72
b 160  75  72   87  75
```

```
i    50
X   60
a    50   60   90
j    50   20   93    80   93
T  160
u    60   50   93  100   88
j    80   50   88    87   97
a    50   20 102    80 121  100  125
s.  210
H  120   33  111
u  100   20  111    50  117   80   121
b  120    8  117   33  121   58   114  83 111
u    60   16  108   66  105
H    70   14  102   57   95  100    90
a  170   11   90   35   95   52   93  64   93   100   93
l   180   16   90   33   88   50   85  66   83    88   83  100   81
a  170   47   74   64   75   82   78 100  78
l   150   40   71   60   74   73   71   93  93
H  320    6   90   15  102   25  102  34 105    46 105    56 102    65   105   75   108
a  170   11  108   29  102   64  102  82   95
d    90   44   86   77   85

u    50   40   85

w    50   20   85   80    83
```

Chaque ligne renferme le phonème à synthétiser, sa durée (en ms) ainsi qu'une série de marqueurs de pitch (peut être aucun), composé chacun de deux nombres entiers : la position du point de pitch à l'intérieur d'un phonème, et la valeur du pitch (en Hz) à cette position. La parole synthétique produite par le synthétiseur MBROLA à partir du fichier de commandes obtenu était de mauvaise qualité. Cela est dû aux erreurs de segmentation faite par le logiciel MBROLIGN. Ce qui nous a ramené à corriger manuellement la segmentation de toutes les phrases du corpus.

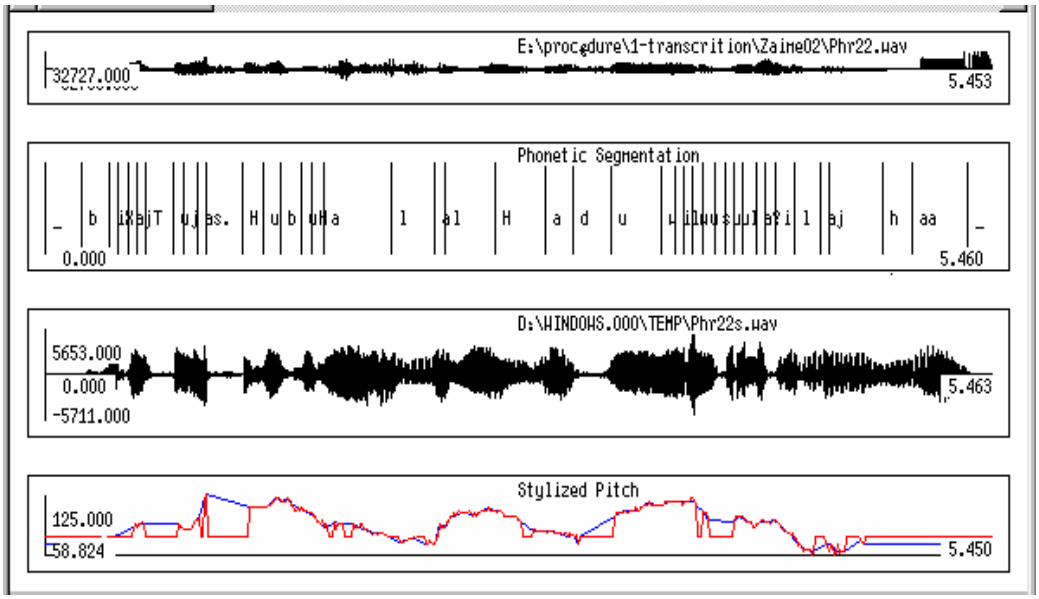


Figure 4.1 : Résultat de segmentation donné par Mbrolign pour la phrase Arabe :

« بحيث يصعب على العدوّ الوصول إليها »

4.4.2. Correction de la segmentation

4.4.2.1. Fonction de variation spectrale SVF

La fonction SVF est un outil de segmentation acoustique. C'est une projection topographique de l'espace multidimensionnel des paramètres spectraux dans un espace unidimensionnel de nombres réels non-négatifs qui refléteraient la mesure de la dissimilitude entre les fenêtres adjacentes d'un signal de parole [140]. De plus nous cherchons une SVF dont les maximums locaux correspondent, autant que possible, avec les changements qualitatifs les plus pertinents de la parole, en particulier, avec les limites des phonèmes.

Nous avons segmenté le signal vocal numérisé en trames de 20 ms avec une fenêtre de décalage de 10ms. Pour chaque trame, un vecteur X de P caractéristiques est calculé.

L'ensemble $S(t,L)=\{X_{t-L},\ldots..,X_{t-1},X_t,X_{t+1},\ldots.X_{t+L}\}$ est une fenêtre SVF de 2L+1 caractéristiques, centrée autour da la trame d'ordre t. Soit S_g et S_d les fenêtres gauche et droite des vecteurs de l'ensemble S(t,L) de part et d'autre du vecteur X_t La Fonction de Variation Spectrale associé à l'ensemble S(t,L) est définie comme étant une projection f_S de cet ensemble en un scalaire a positif :

$$f_S(t): \qquad S(t,L) \xrightarrow{f_s} a \tag{4.1}$$

La fonction f_S doir refléter le degré de dissimilitude entre les vecteurs de S(t,L). Dans notre cas nous avons choisi la distance euclidienne pour mesurer la distance entre deux vecteurs *x* et *y* :

$$dist(x,y)=\sum_{i=1}^{P}(x_i-y_i)^2 \tag{4.2}$$

En utilisant la norme euclidienne, nous devons mesurer la dissimilitude entre la fenêtre gauche S_g et la fenêtre droite S_d par l'expression suivante :

$$f_S(t)=dist(\frac{1}{L}\sum_{i=1}^{L}X_{t-1},\frac{1}{L}\sum_{i=1}^{L}X_{t+1}) \tag{4.3}$$

Cette fonction SVF peut présenter des maximums locaux d'amplitudes très importantes alors que d'autres sont d'amplitudes très faibles. Pour résoudre ce problème, nous devons faire une normalisation en divisant la valeur de la SVF obtenue par la norme au carré du vecteur moyen $\bar{x}$ des vecteurs de l'ensemble S(t,L) :

$$a \to \frac{1}{\left\|\overline{X}\right\|^2} a \tag{4.4}$$

où : $$\overline{X} = \frac{1}{2L+1} \sum_{j=-L}^{j=L} X_j \tag{4.5}$$

Dans notre cas, nous avons pris comme caractéristiques : les 12 premiers coefficients melcepstraux, leurs dérivées dans le temps, leurs dérivées secondes, l'énergie et sa dérivée et sa dérivée seconde. Ces coefficients donnent une bonne représentation de l'enveloppe spectrale locale.

4.4.2.2. Alignement par WAVESURFER

Pour faire une correction manuelle de l'ensemble des phrases du corpus, nous avons utilisé l'outil WAVESURFER développé par le centre CTT (Center of Speech Technology) au KTH à Stockholm. En premier, nous commençons par charger le résultat de segmentation obtenu à partir de MBROLIGN dans WAVESURFER (après une conversion des fichiers phonétiques en fichiers textes contenant dans leurs première et deuxième colonnes : les instants de début et les instants de fin (en secondes) de chaque phonème et une troisième colonne renfermant la liste des phonèmes de la phrase). WAVESURFER nous permet de déplacer les marques (frontières) des phonèmes.

Nous avons calculé également la fonction SVF pour une fenêtre L=2, et chargé la courbe dans WAVESURFER. La correction de la segmentation est basée sur la visualisation de l'onde temporelle, le spectrogramme, et l'écoute des parties sélectionnées du signal, ainsi que les maximums locaux de la fonction SVF.

La figure 4.2 montre le résultat après correction de la segmentation. Cette figure montre de haut en bas : l'onde temporelle, le spectrogramme, la segmentation phonétique, la segmentation en mots et l'évolution temporelle de la fonction SVF. Nous remarquons que la majorité des pics de cette fonction correspondent aux frontières des phonèmes. Donc si des marques sont mal placées, elles seront ajustées par ces maximums locaux.

WAVESURFER nous donne aussi en sortie : un fichier phonétique (.phn) renfermant les phonèmes de la phrase, leurs instants de début et de fin (c'est le fichier phonétique que nous avons chargé au début mais après la correction), un autre fichier (.wrd) renfermant les instants de début et de fin de chaque mot de la phrase, et un fichier contenant l'évolution temporelle de la fréquence fondamentale (.f0).

Pour l'évaluation de la qualité de la segmentation, nous avons construit un fichier de commande (.pho) comme celui obtenu à la sortie de MBROLIGN à partir des fichiers (.phn) et (.f0) d'une même phrase. En synthétisant ce fichier par MBROLA, nous remarquons que la qualité de la parole produite est nettement améliorée, elle est naturelle et intelligible. Ce qui prouve que notre segmentation est bonne.

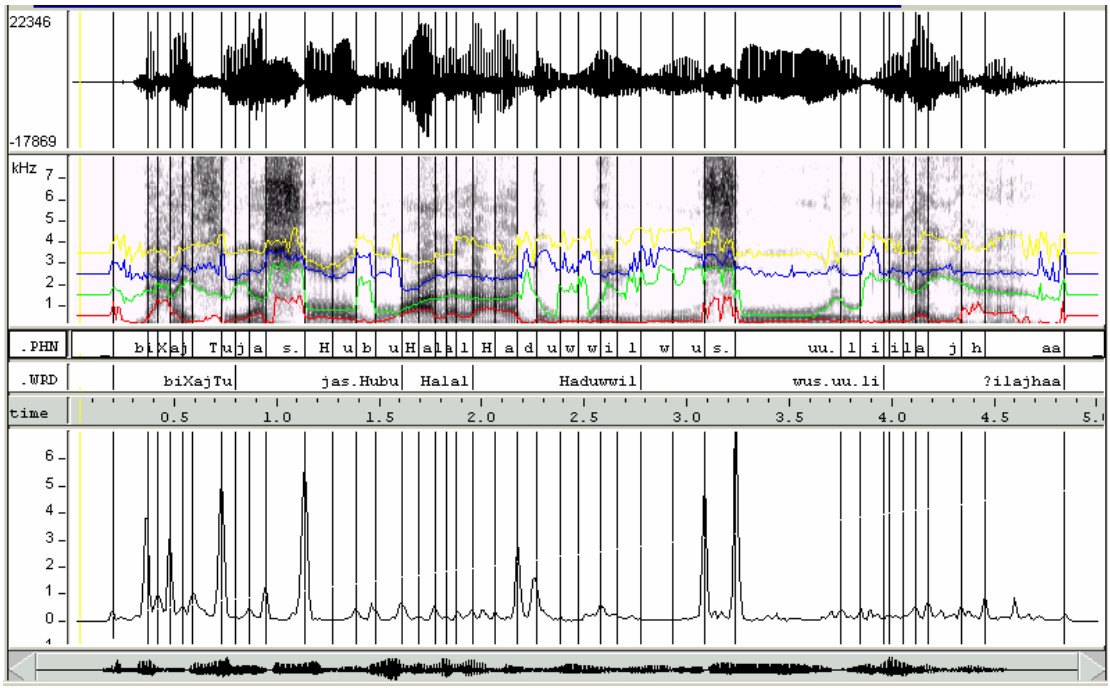


Figure 4.2 : Exemple de segmentation de la phrase Arabe:
" بحيث يصعب على العدوّ الوصول إليها"

4.5. Le fichier de transcription combiné

Nous avons élaboré un programme qui a pour rôle de combiner la transcription en mot et la transcription phonétique pour en faire un seul fichier (.ctf) qui renferme aussi : en plus de la durée du phonème (en ms), la position du phonème dans le mot, le type de syllabe, l'accentuation, ainsi que et le nombre de trames de 10 ms dans chaque phonème.

Une partie du fichier de transcription combiné est donnée dans le tableau 4.2 pour la phrase Arabe : « بحيث يصعب على العدوّ الوصول إليها » :

4.6. Organisation de la base de données

Une base de données est une collection de documents accumulés, qui doivent être extraits sélectivement. Cette propriété rend nécessaire une organisation structurée et hiérarchique de la base de données.

Tableau 4.2 : Exemple de fichier de transcription combiné
Avec AP : Accent Primaire ; AS : Accent secondaire ; Accent Tertiaire

Phonème	Durée (en ms)	Position dans le mot			Type de la syllabe	Accent	Durée (en fen. de 10 ms
_	205	NDebut	NMilieu	NFin	_	AT	20
b	73	Debut	NMilieu	NFin	cv	AT	7
i	47	NDebut	Milieu	NFin	cv	AT	5
X	60	NDebut	Milieu	NFin	cvc	AP	6
a	62	NDebut	Milieu	NFin	cvc	AP	6
j	48	NDebut	Milieu	NFin	cvc	AP	5
T	140	NDebut	Milieu	NFin	cv	AT	14
u	70	NDebut	NMilieu	Fin	cv	AT	7
j	80	Debut	NMilieu	NFin	cvc	AP	8
a	65	NDebut	Milieu	NFin	cvc	AP	6
s.	195	NDebut	Milieu	NFin	cvc	AP	20
H	131	NDebut	Milieu	NFin	cv	AT	13
u	111	NDebut	Milieu	NFin	cv	AT	11
b	93	NDebut	Milieu	NFin	cv	AT	9
u	129	NDebut	NMilieu	Fin	cv	AT	13
H	92	Debut	NMilieu	NFin	cv	AP	9
a	78	NDebut	Milieu	NFin	cv	AP	8
l	57	NDebut	Milieu	NFin	cvc	AS	6
a	49	NDebut	Milieu	NFin	cvc	AS	5
l	76	Debut	NMilieu	Fin	cvc	AS	8
H	112	Debut	NMilieu	NFin	cv	AT	11
a	114	NDebut	Milieu	NFin	cv	AT	11
d	87	NDebut	Milieu	NFin	cvc	AP	9
u	107	NDebut	Milieu	NFin	cvc	AP	11
w	103	NDebut	Milieu	NFin	cvc	AP	10

Notre base de données a été organisée comme la base de sons TIMIT. En plus des fichiers (.wav), six fichiers de transcription (.txt, .trp, .wrd, .phn, .f0, .ctf) existent pour chaque phrase du corpus, qui renferment respectivement:

- le texte de la phrase prononcée (.txt) ;
- la transcription orthographique-phonétique associée au texte (.trp) ;
- la transcription des mots alignée au temps (.wrd) ;
- la transcription phonétique alignée au temps (.phn) ;
- l'évolution temporelle de la fréquence fondamentale (intonation) (.f0) ;
- la transcription combinée.

Les fichiers de transcription phonétique et transcription des mots alignées avec le temps renferment les instants de début et de fin de chaque phonème et chaque mot respectivement. Les fichiers intonatifs (.f0) renferment en plus de l'intonation de la phrase, la décision Voisé/Non-voisé du signal.

4.7. Développement d'un synthétiseur de parole simple utilisant la BD

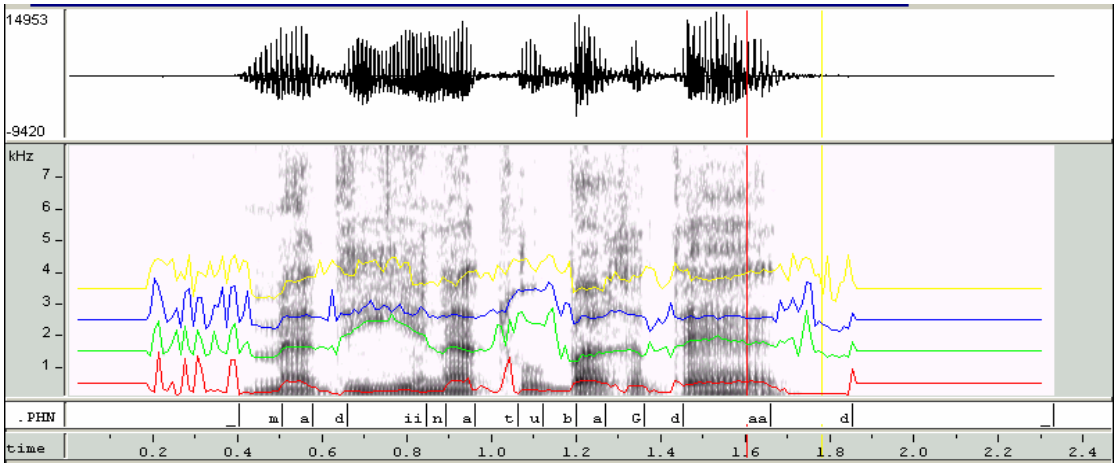
Nous avons créé un synthétiseur vocal simple, basé sur une approche par concaténation de gabarits « templates » pour évaluer les performances de notre base de données prosodiques. Pour se faire, nous avons divisé notre corpus en deux parties : 75% des données pour l'apprentissage et 25% pour le test. Les gabarits sont créés automatiquement à partir des données d'apprentissage. Dans le corpus, chaque phonème peut apparaître plusieurs fois. Pour régler ce problème, il est recommandé de remplacer tous les phonèmes similaires par un seul qu'on appelle gabarits.

Un gabarit est généré pour chaque phonème, en calculant une distance entre chaque paire de phonèmes similaires, et le phonème qui donne la distance minimale entre lui et tous les autres phonèmes similaires sera pris comme gabarit. La mesure de la distance que nous avons choisie est la distance euclidienne entre les coefficients cepstraux qui est une très bonne mesure de la similitude spectrale. Chaque paire de phonèmes subit un alignement temporel par programmation dynamique DTW (Dynamic Time Wrapping) [136]. Ainsi, les phrases tests peuvent être synthétisées à partir de ces gabarits tout en utilisant la méthode TD-PSOLA [2, 67, 68] pour allonger ou réduire le gabarit à la durée désirée pour chaque phonème dans la phrase test.

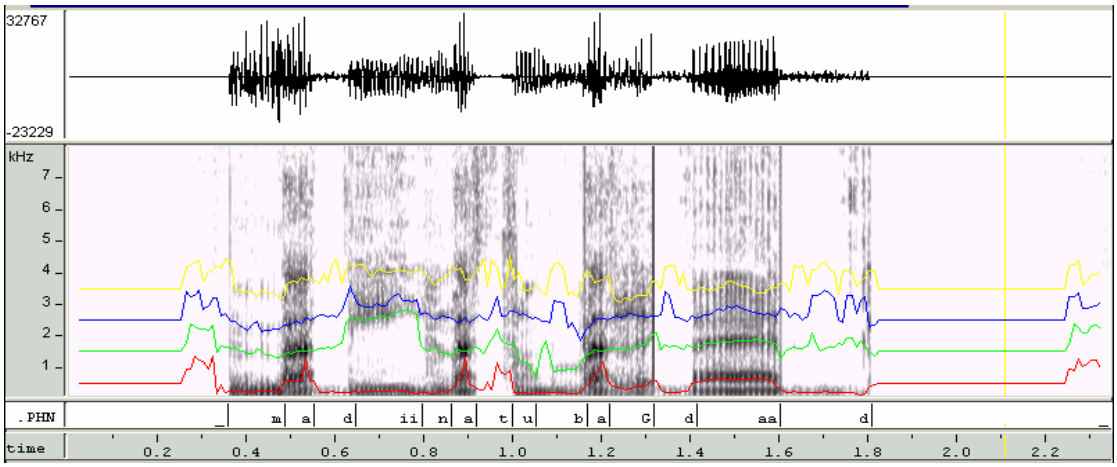
Pour former les gabarits, nous avons essayé les deux cas suivants :

- nous avons pris comme segments similaires de parole : tous les segments associés au même phonème.
- dans ce cas, nous avons pris comme segments similaires : les segments qui ont le même phonème central, même phonème gauche et même phonème droit.

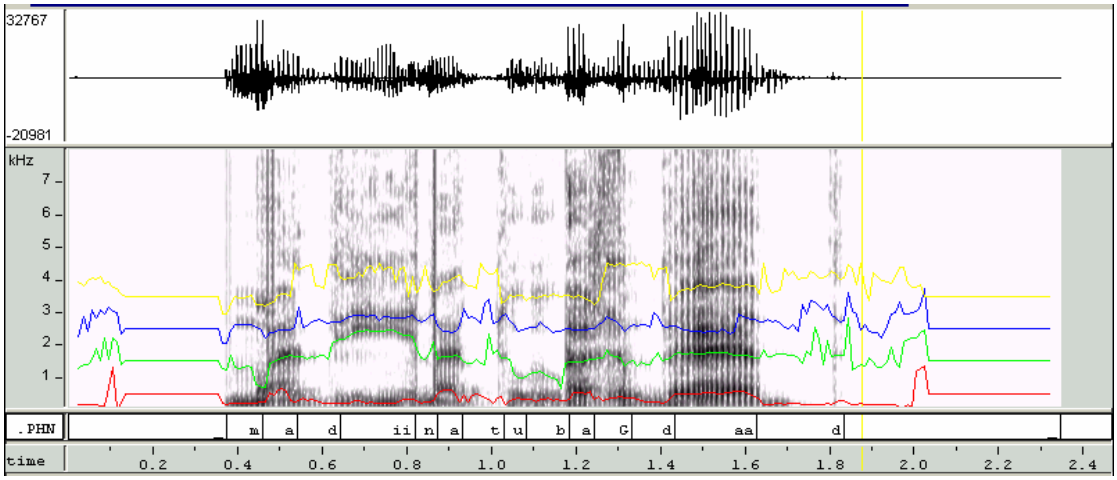
Dans le premier cas, la parole produite par le synthétiseur n'est pas bonne ; elle ne renferme aucune modélisation de la coarticulation entre les phonèmes (c'est une simple concaténation de phonèmes au niveau de leurs zones instables) car chaque gabarit est formées sans tenir compte de son contexte phonétique. Dans le deuxième cas les résultats sont nettement meilleurs, la qualité de la parole produite est intelligible est naturelle (Figure 4.3).



(a)



(b)



(c)

Figure 4.3 : (a) Parole originale, (b) parole synthétique à base de gabarit phonème, (c) parole synthétique à base de gabarit triphone.

4.8. Conclusion

Dans ce chapitre nous venons de présenter les différentes étapes entreprises pour l'élaboration d'une base de données prosodiques de l'Arabe. La segmentation et l'étiquetage du corpus de la parole continue étant la tâche qui demande beaucoup de précision afin d'obtenir une parole synthétique de haute qualité. Au départ, nous avons fait un alignement automatique par l'outil MBROLIGN. Le résultat présentait beaucoup d'erreurs. Nous avons alors corriger ces erreurs manuellement en se basant sur la Fonction de Variation Spectrale SVF dont la majorité des maximums locaux correspondent aux frontières des phonèmes. Pour l'évaluation des résultats obtenus, nous avons utilisé les informations contenues dans les fichiers phonétiques (.phn) et les fichiers intonatifs (.f0) de notre base de données pour créer les commandes du synthétiseur vocal de haute qualité MBROLA. La parole produite est de très bonne qualité. L'application principale d'une telle base de données est la génération automatique de la prosodie pour l'Arabe, et la réalisation d'un synthétiseur vocal par réseau neuronal.

CHAPITRE 5
UTILISATION DES RESEAUX DE NEURONES POUR LA MODELISATION DE LA PROSODIE ET LA GENERATION DES PARAMETRES DE SYNTHESE D'UN CODEUR RELP

5.1. Introduction

Ce chapitre décrit les étapes de développement d'un système de synthèse TTS Arabe à base de Réseaux de Neurones et d'un modèle par Prédiction Linéaire à Excitation Résiduelle (RELP). Les notions de base des RN sont présentées dans la première partie du chapitre. Dans le reste du chapitre nous mettons l'accent sur l'implémentation d'un synthétiseur de haute qualité. Dans ce travail, les réseaux RN sont utilisés pour la génération des paramètres de contrôle d'un codeur par prédiction linéaire LPC et la synthèse des informations prosodiques. Une évaluation subjective a été effectuée pour montrer les performances de notre synthétiseur en comparaison avec un synthétiseur par sélection d'unités que nous avons également réalisé. Les résultats et discussions sont donnés dans la dernière partie du chapitre.

5.2. Notions sur les Réseaux de Neurones Artificiels (RNA)

L'intelligence artificielle est en continuelle progression depuis l'invention de l'ordinateur et l'utilisation de programmes informatiques. Il existe en effet de nombreux programmes capables de réaliser des choses de plus en plus complexes : diriger un robot, résoudre des problèmes, jouer aux échecs, etc. Mais ils sont très rarement capables de rivaliser avec le cerveau humain. En effet, les ordinateurs n'ont pas la faculté d'apprentissage (qui est une caractéristique humaine), ils ne connaissent pas le progrès si personne ne les modifie, et c'est pour cela que de nombreuses tâches sont encore irréalisables par les ordinateurs.

D'un autre côté, la neurobiologie a apporté un grand nombre d'informations sur le fonctionnement du système nerveux (cerveau, neurones biologiques, etc.). Des mathématiciens ont alors tenté de reproduire le fonctionnement du cerveau en intégrant ces

connaissances en biologie dans des programmes informatiques, et en leur donnant la possibilité d'apprendre.

Inspirés des réseaux neuromémitiques, plusieurs modèles de RNA existent, et chaque modèle se prête bien pour une application particulière (classification, synthèse, reconnaissance, contrôle, etc.). Cependant, malgré tous les efforts déployés en algorithmique et en intelligence artificielle, leurs utilisations restent limitées dans quelques applications, et il reste beaucoup de domaines où les RN n'ont pas trouvé de solutions. Les meilleurs systèmes à réseaux de neurones restent assez loin d'imiter des performances telles que celles de l'être humain.

5.2.1 Le système neuronal physiologique

L'unité cellulaire de base du système nerveux est le neurone. Le cerveau se compose d'environ 10^{11} neurones, et chacun est connecté à environ 10.000 autres neurones. Les connexions permettent le transfert d'informations sous forme d'impulsions électriques entre les neurones.

Un neurone biologique est constitué d'un corps cellulaire et d'un noyau. Il se ramifie de dendrites par où s'achemine l'information vers le neurone. Une fois l'information traitée par le neurone elle est envoyée sur l'axone qui la propage vers d'autres cellules. Les contacts entre deux neurones, de l'axone à une dendrite, se font par l'intermédiaire des synapses. Les informations parvenant au neurone peuvent venir d'autres neurones ou de capteurs sensoriels tels que le touché, la vue, l'odorat, etc., ou d'autres capteurs internes à l'organisme (Figure 5.1).

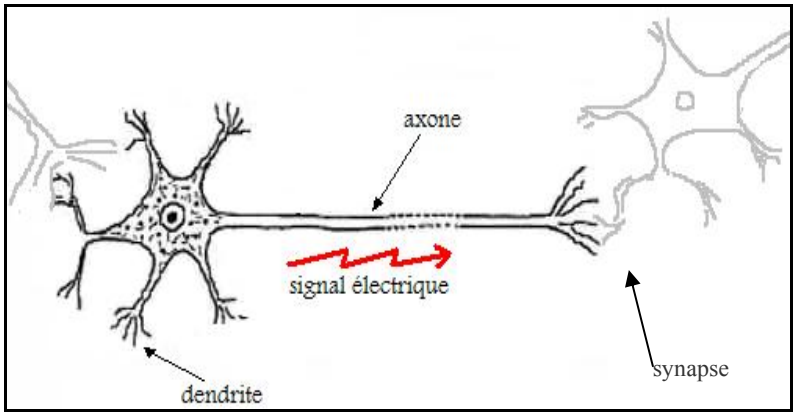


Figure 5.1 : Réseau de Neurones biologique

5.2.2. La modélisation mathématique du système biologique

Formellement, un réseau de neurones est un graphe dont les noeuds sont des unités de calcul appelées neurones formels, et les arrêtes orientées et pondérées sont appelées liens synaptiques (figure 5.2). Chaque neurone artificiel est un processeur élémentaire. Il reçoit un certain nombre d'entrées en provenance de neurones en amonts ou des capteurs composant la machine dont il fait partie. Chacune de ces entrées est pondérée par un poids synaptique représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un certain nombre de neurones en aval.

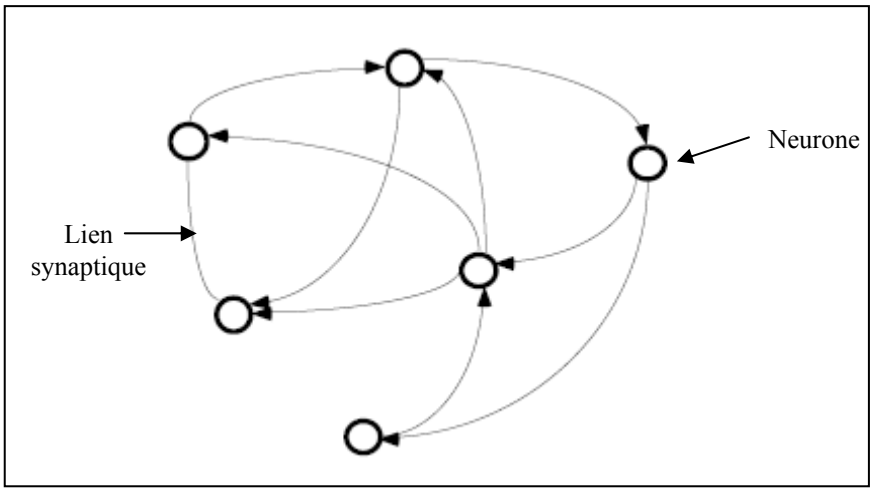


Figure 5.2 : Modèle d'un réseau de neurones

Un neurone formel effectue tout d'abord une somme des valeurs d'entrée pondérées (Figure 5.3). Cette somme sera ensuite modifiée par une fonction de transfert qui permet d'obtenir la valeur effective de l'activation de la cellule, valeur qui sera répercutée en sortie. Il existe de nombreuses formes possibles pour la fonction de transfert. Les plus courantes sont présentées sur les figures 5.4, 5.5 et 5.5.

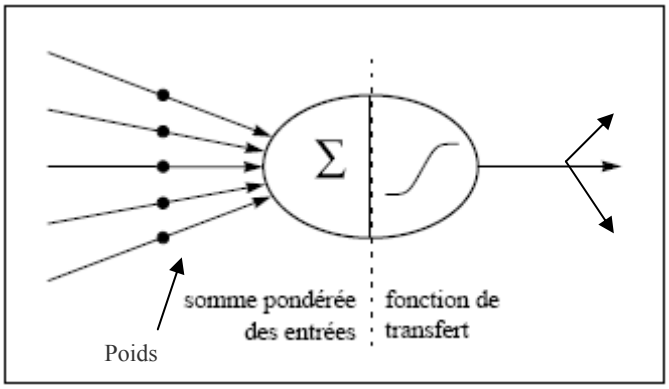


Figure 5.3 : Neurone formel

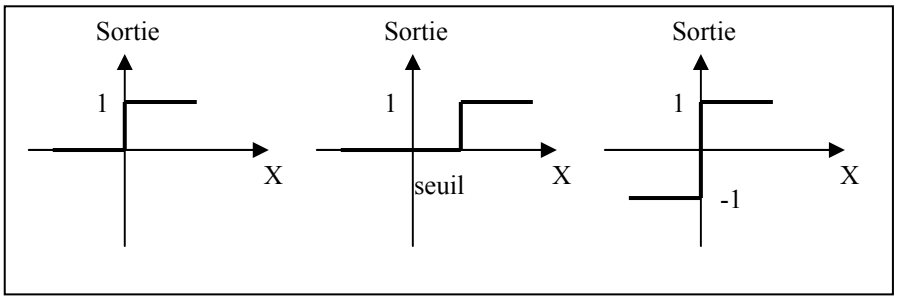


Figure 5.4 : Exemples de fonctions binaires (X est la somme des entrées)

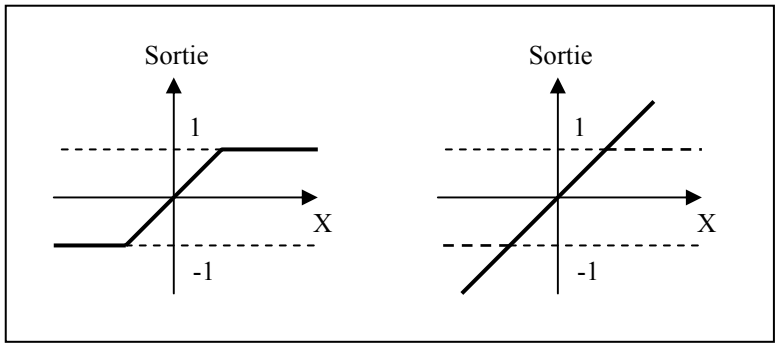


Figure 5.5 **:** Exemples de fonctions linéaires
(a) : fonction linéaire à saturation ; (b) : fonction linéaire

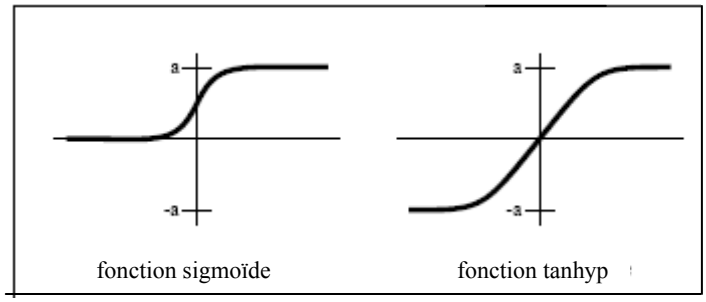


Figure 5.6 : Exemples de fonctions non linéaires dérivables
(a) : fonction sigmoïde ; (b) : fonction tangeante hyperbolique

5.2.3. Architectures des Réseaux de Neurones

Il est possible de distinguer deux types d'architecture : les modèles statiques et les modèles dynamiques.

5.2.3.1. Réseaux de neurones statiques

Les modèles statiques sont actuellement les plus utilisés. La principale caractéristique d'un modèle statique est qu'il permet de classer des formes indépendantes du temps et

d'une quelconque évolution. La forme à classer à un instant *t* est donc jugée totalement indépendante des formes classées lors des instants précédents. Les réseaux statiques se caractérisent donc avant tout par l'absence totale de récurrence au sein de leur architecture [141]. Ce type de réseaux a connu d'importants développements et de nombreuses architectures ont été définies pour résoudre des problèmes variés. Nous citons les réseaux perceptrons multicouches qui sont largement utilisés à cause de leur efficacité et leur facilité d'implémentation. Leur architecture et leur méthode d'apprentissage seront détaillées ultérieurement car, c'est le type de réseau que nous avons utilisé dans notre travail. Nous citons également les réseaux de neurones RBF (Radial Basis Function) et les modèles à auto-organisation tels que la carte de KOHONEN [142] qui est principalement utilisée en reconnaissance de la parole.

5.2.3.2. Réseaux de neurones dynamiques

La caractéristique distinctive des réseaux dynamiques (appelés encore réseaux récurrents) par rapport aux réseaux statiques est la mise en place de connexions récurrentes dans l'architecture. Ces réseaux sont organisés tel que chaque neurone reçoit sur ses entrées une partie ou la totalité de l'état du réseau (sortie des autres neurones) en plus des informations externes. Ainsi, l'état d'un neurone est fonction du vecteur d'entrée courant et de l'état du réseau à l'instant précédent. A titre d'exemples, nous citons les modèles de HOPFIELD, de JORDAN et d'ELMAN, ainsi que les réseaux multicouches avec temps de retard TDNN (Time Delay Neural Network).

5.2.4. Apprentissage

Les réseaux de neurones artificiels possèdent la faculté d'apprentissage (par exemple à reconnaître une lettre, un son...). Dans la majorité des algorithmes actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions (et éventuellement d'autres paramètres, notamment les seuils). L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience. Il est souvent impossible de décider à priori les valeurs des poids des connexions d'un réseau pour une application donnée. A l'issue de l'apprentissage, les poids sont fixés : c'est alors la phase d'utilisation.

Les réseaux de neurones se divisent en deux principales classes, les réseaux à apprentissage supervisés (Supervised Learning) et les réseaux à apprentissage non supervisés (Unsupervised Learning). Dans le cas d'un apprentissage supervisé, on présente

au réseau les entrées et les sorties désirées correspondantes. Le réseau doit alors se reconfigurer, c'est-à-dire calculer ses poids de façon à réduire l'erreur entre la sortie qu'il donne et la sortie désirée.

L'apprentissage dans les modèles non supervisés, comme le réseau de KOHONEN, se fait grâce à l'emploi d'une fonction de voisinage. Après initialisation aléatoire des poids, une forme est présentée en entrée du réseau. L'apprentissage entre alors dans une phase de compétition : c'est la cellule dont le potentiel d'activation est le plus fort en fonction de l'entrée qui est choisie comme vainqueur. Cette activation est calculée en fonction d'une distance et sera d'autant plus forte que cette distance entre les poids synaptiques de la cellule et les valeurs du vecteur d'entrée sera faible. Le choix d'une cellule particulière permet alors d'ajuster les poids localement, en minimisant la différence qui existe encore entre les poids et le vecteur d'entrée. Cet ajustement se fait suivant une forme de voisinage qui peut être carrée, ronde ou hexagonale. La taille du voisinage décroît de manière progressive lors de l'apprentissage et les valeurs des connexions sont ajustées selon une certaine fonction de voisinage [143].

L'apprentissage neuromémitique (supervisé ou non supervisé) nécessite en général une grande quantité de données, que l'on regroupe dans ce que l'on appelle « corpus d'apprentissage ». Selon la technique d'apprentissage utilisée, d'autres corpus sont aussi employés, notamment pour mesurer la validité de la solution trouvée par le réseau. On appelle ces corpus supplémentaires, des corpus de test ou de généralisation.

5.2.5. Les perceptrons multicouches

5.2.5.1. Architecture

Un Perceptron MultiCouche PMC (en Anglais Multi Layer Perceptrons : MLP) comprend une couche d'entrée qui correspond aux variables d'entrée, une couche de sortie, et un certain nombre de couches intermédiaires appelées couches cachées. Les liens n'existent qu'entre les cellules d'une couche avec les cellules de la couche suivante (figure 5.7).

Plusieurs couches de traitement permettent au PMC de réaliser des associations non linéaires entre l'entrée et la sortie. Ce qui permet à ce type de réseau d'aborder des problèmes plus difficiles. Les PMC sont largement utilisés dans les problèmes de classification, de traitement d'images ou de diagnostic.

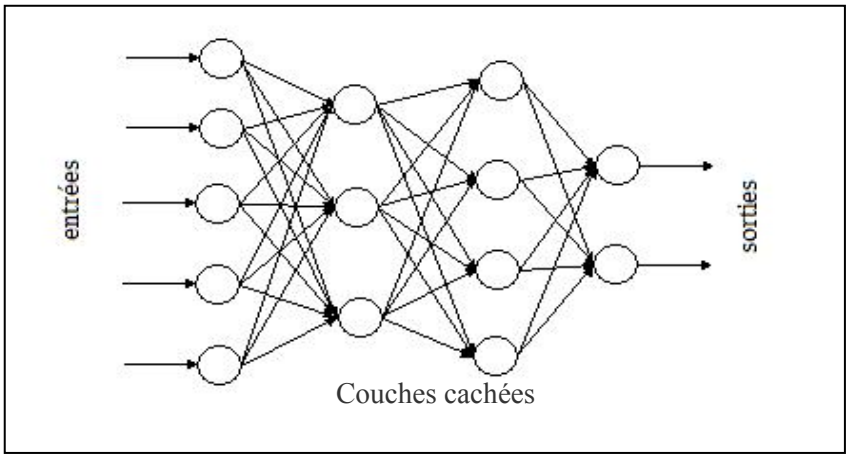


Figure 5.7 : Réseau multicouche

5.2.5.2. Apprentissage des réseaux perceptrons multicouches

L'apprentissage dans les réseaux connexionnistes du type des perceptrons multicouches est un apprentissage supervisé qui se fait grâce à la méthode de la rétropropagation du gradient d'erreur [144,145]. Cette méthode est largement utilisée pour calculer le gradient d'une fonction d'erreur. Elle consiste à effectuer le calcul des dérivées partielles de la fonction neuronale vis à vis des paramètres du réseau en partant de la couche de sortie, puis de se servir de ces quantités pour calculer les dérivées partielles dans la couche cachée, avant de remonter vers la couche d'entrée. Cette méthode présente l'avantage d'être économique en termes de nombre d'opérations à effectuer.

Soit un PMC défini par une architecture à *n* entrées et *p* sorties. L'erreur du PMC sur un échantillon d'apprentissage S d'exemples $(\vec{x}^s, \vec{t}^s)$ est définie par :

$$E = \frac{1}{2} \sum_{(\vec{x}^s, \vec{t}^s) \in S} \sum_{k=1}^{p} (t_k^s - o_k^s)^2$$

où o_k^s est la k^{ème} composante du vecteur de sortie $\vec{o}^s$ calculée par le PMC pour une entrée $\vec{x}^s$; t_k^s étant la k^{ème} composante du vecteur de sortie désirée pour la même entrée. Cependant, au lieu de chercher à minimiser l'erreur globale sur l'échantillon complet, la méthode de rétropropagation du gradient tend à minimiser l'erreur sur chaque présentation individuelle d'exemple. Ainsi, l'erreur de sortie du réseau pour un exemple du corpus d'apprentissage est calculée comme suit (en imaginant que l'ensemble d'apprentissage s'est réduit à un seul exemple) :

$$E = \frac{1}{2}\sum_{k=1}^{p}(t_k - o_k)^2 \tag{5.1}$$

La rétropropagation du gradient consiste à déterminer les poids synaptiques du réseau neuronal de façon à minimiser l'erreur quadratique *E* sur tous les neurones de sortie. Pour cela, la méthode de descente de gradient est utilisée. Elle permet de modifier les poids w_{ij} d'une quantité Δw_{ij} :

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{5.2}$$

où :

- $0 \leq \eta \leq 1$: le taux d'apprentissage ;
- $\vec{x}_i$: le vecteur d'entrée pour le neurone i (x_{ij} étant la j^{ème} entrée du neurone i) ;
- $\vec{w}_i$: le vecteur des poids synaptiques pour le neurone i (w_{ij} est le poids associé au lien entre le neurone j vers le neurone i).

Considérons aussi les notations suivantes (figure 5.8) :

- *Pred(i)* : l'ensemble des cellules dont la sortie est une entrée de la cellule i ;
- *Succ(i)* : l'ensemble des cellules qui prennent comme entrée la sortie de la cellule i ;
- $z_i = \vec{w}_i . \vec{x}_i = \sum_{j \in Pred(i)} w_{ij} x_{ij}$, la somme pondérée des entrées du neurone i ;
- o_i est la sortie de la cellule i ($o_i = f(z_i)$, f étant la fonction d'activation du neurone i).

Nous pouvons écrire alors :

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}} = \frac{\partial E}{\partial z_i} x_{ij} \tag{5.3}$$

Pour simplifier quelques notations, nous appelons δ_i la quantité $-\partial E / \partial z_i$. Cette valeur se calcule différemment en fonction de la position du neurone dans l'architecture. Deux cas doivent être considérés : soit le neurone est une des unités de la couche de sortie, soit le neurone appartient à une des couches cachées du réseau.

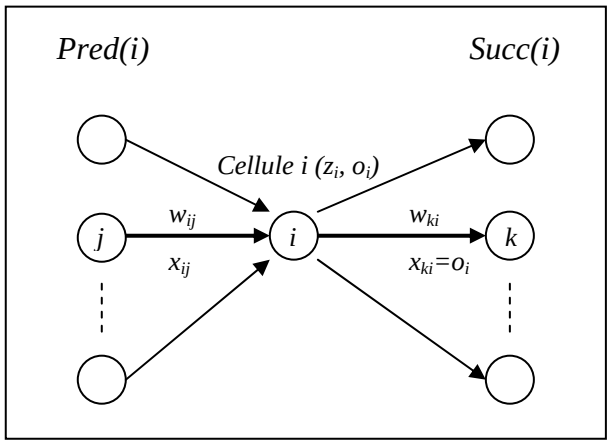


Figure 5.8 : Notations utilisées

Equation pour un neurone de la couche de sortie

Dans ce cas, la quantité z_i (la somme pondérée des entrées du neurone i) ne peut influencer la sortie du réseau que par le calcul de o_i. Nous avons donc :

$$\delta_i = -\frac{\partial E}{\partial z_i} = -\frac{\partial E}{\partial o_i} \cdot \frac{\partial o_i}{\partial z_i} \tag{5.4}$$

La première dérivée partielle est donnée par :

$$\frac{\partial E}{\partial o_i} = \frac{\partial}{\partial o_i}\left(\frac{1}{2}\sum_{k=1}^{p}(t_k - o_k)^2\right)$$

Seul le terme correspondant à $k=i$ a une dérivée non nulle, ce qui donne :

$$\frac{\partial E}{\partial o_i} = \frac{\partial}{\partial o_i}\frac{1}{2}(t_i - o_i)^2 = -(t_i - o_i) \tag{5.5}$$

Pour la seconde dérivée partielle de l'équation (5.4) nous avons :

$$\frac{\partial o_i}{\partial z_i} = \frac{\partial}{\partial z_i} f(z_i) = f'(z_i) \tag{5.6}$$

où f est la fonction d'activation de la couche de sortie.

Dans le cas où la fonction d'activation utilisée est une fonction sigmoïde dérivable définie par :

$$f(x) = \frac{1}{1+e^{-x}}$$

Alors l'équation (5.6) devient :

$$\begin{aligned}\frac{\partial o_i}{\partial z_i} &= \frac{\partial}{\partial z_i}\left(\frac{1}{1+e^{-z_i}}\right) = \frac{e^{-z_i}}{(1+e^{-z_i})^2} \\ &= o_i\left(\frac{e^{-z_i}}{1+e^{-z_i}}\right) = o_i\left(\frac{1+e^{-z_i}}{1+e^{-z_i}} - \frac{1}{1+e^{-z_i}}\right) \\ &= o_i(1-o_i)\end{aligned} \tag{5.7}$$

En substituant les résultats obtenus par les équations (5.5) et (5.7) dans l'équation (5.4), nous obtenons :

$$\delta_i = -\frac{\partial E}{\partial z_i} = f'(z_i)(t_i - o_i) = o_i(1-o_i)(t_i - o_i) \tag{5.8}$$

et par conséquent l'équation (5.2) devient :

$$\Delta w_{ij} = \eta \delta_i x_{ij} \tag{5.9}$$

Equation pour un neurone d'une couche cachée

Dans le cas où le neurone *i* est en couche cachée, la valeur de l'erreur relative à ce neurone doit être calculée en fonction de l'ensemble des erreurs effectuées par les neurones auxquels le neurone considéré est connecté en aval (c'est-à-dire les neurones de l'ensemble *Succ(i)*). Nous avons alors :

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}} = -\delta_i \cdot x_{ij} = \sum_{k \in Succ(i)} \frac{\partial E}{\partial z_k} \cdot \frac{\partial z_k}{\partial o_i} \cdot \frac{\partial o_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ij}} \tag{5.10}$$

Notons que :

$$-\frac{\partial E}{\partial z_k} = \delta_k, \quad \frac{\partial z_k}{\partial o_i} = w_{ki}, \quad \frac{\partial o_i}{\partial z_i} = o_i(1-o_i) \text{ et } \frac{\partial z_i}{\partial w_{ij}} = x_{ij}$$

D'où :

$$\frac{\partial E}{\partial w_{ij}} = o_i(1-o_i)x_{ij} \sum_{k \in Succ(i)} \frac{\partial E}{\partial z_k} \cdot w_{ki} = -o_i(1-o_i)x_{ij} \sum_{k \in Succ(i)} \delta_k \cdot w_{ki} \tag{5.11}$$

et

$$\delta_i = o_i(1-o_i) \sum_{k \in Succ(i)} \delta_k \cdot w_{ki} \tag{5.12}$$

La modification du poids w_{ij} est alors définie par :

$$\Delta w_{ij} = \eta x_{ij} \delta_i$$

Maintenant, nous sommes maintenant en mesure pour énoncer formellement l'algorithme de rétropropagation du gradient.

Algorithme de rétropropagation du gradient :

- Entrée : Un échantillon S de $R^n \times R^p$; η ; un réseau PMC avec une couche d'entrée, L-1 couches cachées, et une couche de sortie.
- Initialisation de tous les poids à de petites valeurs aléatoires dans l'intervalle [-0.5, 0.5]
- Jusqu'à ce que le critère d'arrêt soit vérifié, Faire :
  - Pour chaque exemple d'apprentissage $(\vec{x}, \vec{t})$ de S, Faire :
    1. Calculer la sortie $\vec{o}$ du PMC correspondant à l'entrée $\vec{x}$,
    2. Pour chaque neurone de sortie i calculer :

$$\delta_i = f'(z_i)(t_i - o_i) = o_i(1-o_i)(t_i - o_i)$$

    3. Pour chaque couche de L-1 à 1,

       Pour chaque cellule *i* de la couche courante calculer :

$$\delta_i = f'(z_i) \sum_{k \in Succ(i)} \delta_k \cdot w_{ki} = o_i(1-o_i) \sum_{k \in Succ(i)} \delta_k \cdot w_{ki}$$

    4. Mise à jour des poids w_{ij} comme suit :

$$w_{ij} \leftarrow w_{ij} + \eta \delta_i x_{ij}$$

- Sortie : un PMC défini par la structure initiale choisie et les nouveaux poids w_{ij}

Rappelons que la méthode de rétropropagation du gradient d'erreur exige un bon choix du taux d'apprentissage η pour fonctionner correctement : si η est trop petit la convergence vers une solution optimale est lente. S'il est trop grand, on risque d'osciller autour du minimum. Dans certain cas on risque aussi d'avoir des problèmes de minima locaux. Pour pallier à ces problèmes, une solution consiste à pondérer la modification des poids en fonction du nombre d'itérations déjà effectué. La règle de modification des poids à instant donné t devient :

$$\Delta w_{ij}(t) = \eta x_{ij} \delta_i + m \Delta w_{ij}(t-1)$$

où $0 \leq m < 1$ est un coefficient appelé momentum.

5.3. Description de notre système TTS à base de réseaux de neurones

La figure 5.9 montre le système de synthèse à partir du texte TTS (Text-To-Speech) Complet. Ce dernier inclue un sous-système pour le traitement linguistique, trois réseaux de neurones (RN) utilisés pour déterminer les informations prosodiques (durée, gain, et intonation) de chaque segment phonétique. Un autre réseau neuronal est utilisé pour

convertir la description linguistique du texte et les durées segmentales en une série de paramètres d'un codeur à prédiction linéaire LPC. Ce système renferme aussi la partie synthèse d'un codeur de parole paramétrique qui utilise un modèle source-filtre. Ce filtre est décrit par les coefficients LSF (Line Spectral Frequencies). La source de ce codeur est obtenue à partir d'un dictionnaire d'excitation. Les segments résiduels extraits de ce dictionnaire vont subir une modification de la durée et de l'intonation, une concaténation et une multiplication de gain. Le codeur pourra ainsi convertir les paramètres LSF en signal parole.

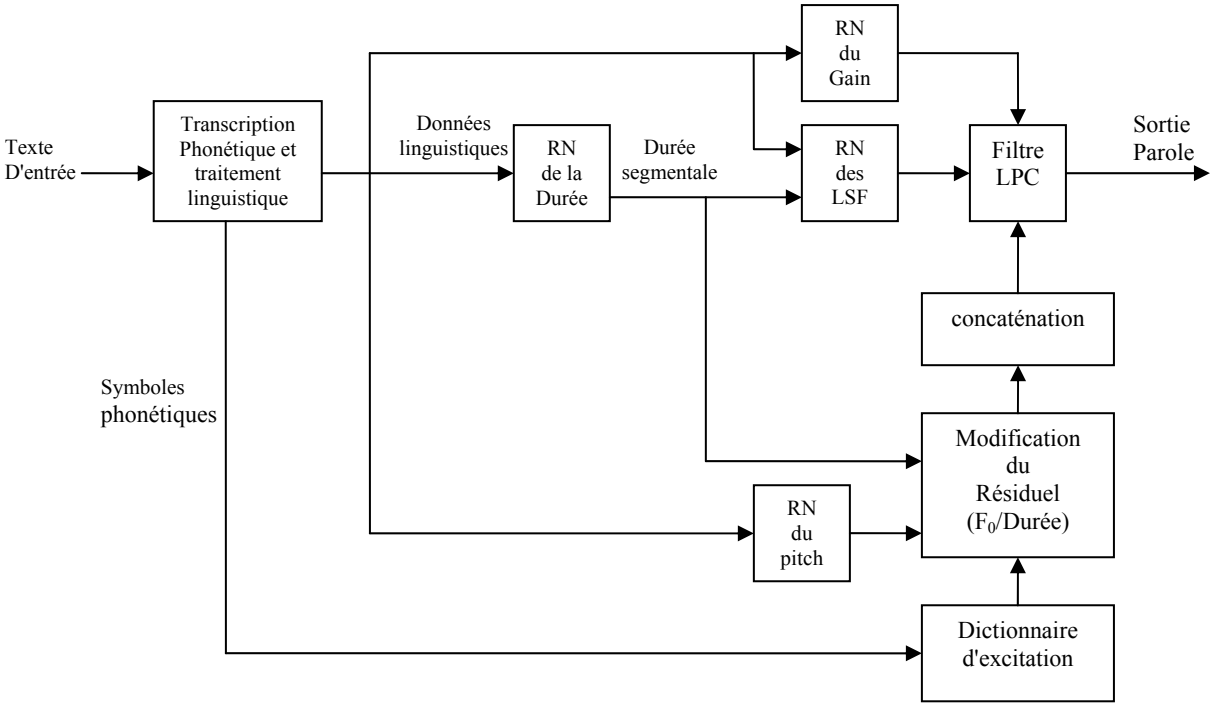


Figure 5.9 : Système TTS à base de réseaux de neurones

5.4. Données d'apprentissage, de validation et de test

Dans l'ordre de faire l'apprentissage d'un réseau de neurone pour effectuer une bonne projection phonétique-acoustique, nous avons utilisé la base de données prosodique (voir chapitre 4). Rappelons que cette base de données renferme l'enregistrement d'un ensemble de phrases arabes prononcées par un seul locuteur. Ces dernières sont ensuite étiquetées phonétiquement, prosodiquement, et syntaxiquement. Les enregistrements sont ensuite analysés par le codeur LPC pour produire des vecteurs de paramètres LPC décrivant chacun les caractéristiques acoustiques d'une fenêtre de parole égale à 10 ms. Les étiquettes de parole sont aussi traitées pour générer les vecteurs d'entrées du réseau

neuronal, décrivant les contextes phonétiques et prosodiques des fenêtres de 10 ms de parole. Les différentes étapes pour la génération des données d'apprentissage seront décrites en détail dans les paragraphes suivants.

5.4.1. Caractéristiques Articulatoires de l'Arabe et codage des entrées ANN

La représentation de l'information dans une couche d'entrée d'un réseau de neurones est très importante. La manière avec laquelle l'entrée est codée influence beaucoup les performances du réseau. En plus, le réseau doit être capable de faire la généralisation pour de nouvelles situations. Par conséquent la structure de l'entrée doit refléter facilement les ressemblances entre les classes d'entrée. Dans ce travail, nous désirons que notre réseau produise les coefficients LSF pour chaque fenêtre de synthèse de 10 ms. L'entrée qui permet de produire chaque fenêtre de sortie inclut le contexte, la position, l'accent, la durée du phonème, et la position de la fenêtre à l'intérieur du phonème courant. Un schéma de codage basé sur les caractéristiques articulatoires et lexicales du phonème à synthétiser peut accomplir cet objectif. Donc chaque phonème sera représenté par un vecteur binaire, traduisant sa transcription en traits distinctifs donnés par le tableau 5.1**.** Les vecteurs sont de dimension 28 bits qui représentent le nombre de caractéristiques prises en considération pour la transcription. Si le phonème montre une caractéristique donnée, sa position relative dans le vecteur sera marquée par un 1, sinon elle sera marquée par un 0.

Tableau 5.1 : Les différentes caractéristiques articulatoires et lexicales des phonèmes Arabes

Selon le voisement	**Type du phonème**	**Autres caractéristiques**
• Voisé/Non voisé	• Plosive	• Emphatique
Lieu d'articulation	• Fricative	• Pause
• Bilabial	• Nasale	**Position dans le mot**
• Labiodental	• Latérale	• Début
• Interdental	• vibrante	• Milieu
• Alvéo-dental	• Voyelle	• Fin
• Palatale	• Semi voyelle	**Accentuation**
• Vélaire	**Type de voyelle**	• Accent primaire
• Uvulaire	• Antérieur/Postérieur	• Accent secondaire
• Pharyngale	• Ouvert/Fermé	• Accent tertiaire (inaccentué)
• laryngale	• Court/Long	

Par exemple : le tableau suivant donne les codes des phonèmes /m/, /a/, et /d/ selon leurs caractéristiques articulatoires et lexicales dans la phrase Arabe « **مدينة بغداد** »

Tableau 5.2 : Exemple de codage

	Voisé/Nonvoisé	Plosive	Fricative	Nasale	Vibrante	Latérale	Semi-voyelle	Longue/Courte	Antérieur/Postérieur	Férmé/Ouvert	Bilabiale	Labio-dentale	Dentale	Alvéo-dentale	Palatale	Vélaire	Uvélaire	Pharyngale	Laryngale	Emphatique	Voyelle	Pause	Début	Milieu	fin	Accent primaire	Accent secondaire	Accent tertiaire
m	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
a	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
d	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0

5.4.1. Codeur LPC et expansion de la largeur de Bande

Notre réseau neuronal doit associer à chaque représentation phonétique en entrée, les paramètres acoustiques correspondants. Ces paramètres sont générés par un codeur à prédiction linéaire LPC. Ce codeur analyse les signaux des fichiers wav du corpus sur des fenêtres de Hamming de 30 ms, avec un pas fixe de 10 ms. 10 coefficients LSF sont calculés pour chaque pas. Ces informations obtenues par analyse LPC constituent les sorties désirées de notre réseau de neurones. Nous avons choisi la représentation LSF des coefficients LPC car elle possède d'excellentes propriétés d'interpolation et de quantization pour le codage à bas débit [14, 27, 140]. Il a été démontré que ces propriétés sont très utiles dans l'apprentissage des réseaux de neurones et qu'elles donnent de meilleurs résultats par rapport à ceux obtenus en utilisant les autres représentations [15]. Les coefficients LSF sont écrits comme suit :

$$0 = LSF_0 < LSF_1 < LSF_2 < \cdots < LSF_p < LSF_{p+1} = \pi \qquad (5.13)$$

La condition ci-dessus assure la stabilité du filtre de synthèse [145].

Occasionnellement, l'analyse LPC génère un filtre de synthèse avec des crêtes spectrales de formants pointues. Ceci implique que les pôles du filtre sont très proches du cercle unité et par conséquent, le filtre est marginalement stable. Une telle stabilité marginale dans les filtres LP peut causer des gazouillements occasionnels dans la parole synthétique. Pour pallier ces problèmes, une solution est d'utiliser l'expansion de largeur de bande (utilisée à l'origine dans le codage de la parole) pour augmenter les largeurs de bande de la réponse en fréquence du filtre [147].

Chaque coefficient a_k est remplacé par $\gamma^k a_k$, où $k=1,2, \ldots, p$ (ceci est équivalent à l'usage du filtre d'analyse $A'(Z) = A(\gamma Z)$, où $A(Z)$ est le filtre LPC). De telle multiplication

déplace tous les pôles du filtre du cercle unité vers l'origine avec un facteur γ . Le résultat en est donc des crêtes lissées et des bandes de fréquences élargies dans la réponse en fréquence du filtre d'analyse et ainsi, le filtre devient plus stable. Dans notre cas, nous avons choisi un facteur d'expansion de largeur de bande γ égal à 0.996 qui correspond à un élargissement de 30 Hz. Après ceci, les coefficients LP modifiés sont convertis en coefficients LSF. Ces derniers constituent les sorties désirées qu'on utilise pour l'apprentissage de notre réseau de neurones.

5.4. Réseau de neurone phonétique-acoustique

L'objectif de ce réseau neuronal est de générer les paramètres LSF du filtre de synthèse. L'architecture et l'apprentissage de ce réseau, ainsi que le lissage spectral des paramètres LSF aux points de concaténation entre phonèmes adjacents sont décrits dans cette section.

5.5.1. Architecture

Dans ce travail, nous avons utilisé le Perceptron Multi Couches (PMC), connu comme étant le type de réseaux le plus répandu et le plus utilisé, vu la simplicité de sa structure et la rapidité de son apprentissage (voir la section 5.2). Le PMC que nous avons utilisé renferme une seule couche cachée avec 10 neurones. Son architecture est montrée sur la figure 5.10. La couche d'entrée forme une fenêtre glissante sur le jet de données à l'entrée. Elle contient trois groupes de neurones, qui représentent le phonème courant, le phonème du contexte gauche, et celui du contexte droit. Chacun des phonèmes est représenté par un vecteur de caractéristiques articulatoires et lexicales. En plus de ces trois groupes de neurones, dix neurones index de temps sont utilisés pour indiquer la variation temporelle à l'intérieur du phonème courant [15], ils représentent la position relative de la fenêtre courante dans le phonème. Ceci aide au lissage des transitions entre les vecteurs voisins essentiellement aux frontières des segments. La valeur à la sortie du neurone index i (i allant de 1 à 10) durant la fenêtre j est calculée en utilisant l'équation (5.14) (nous avons choisi β =0.2), telle que la sortie de i atteint sa valeur maximale durant la fenêtre j=i. Un autre neurone d'entrée représente la durée normalisée du phonème courant.

$$O_i = \exp(-\beta(i-j)^2) \tag{5.14}$$

Le nombre de neurones dans la couche d'entrée étant donc égale à 95. La couche de sortie renferme 10 neurones qui représentent les paramètres du codeur LPC (10

coefficients LPC). Dans notre cas, nous avons choisi les coefficients LSF pour représenter les paramètres du codeur LPC car ils nous donnent l'erreur d'apprentissage la plus faible. Les fonctions d'activation utilisées sont la fonction sigmoïde dans la couche cachée, et la fonction linéaire dans la couche de sortie.

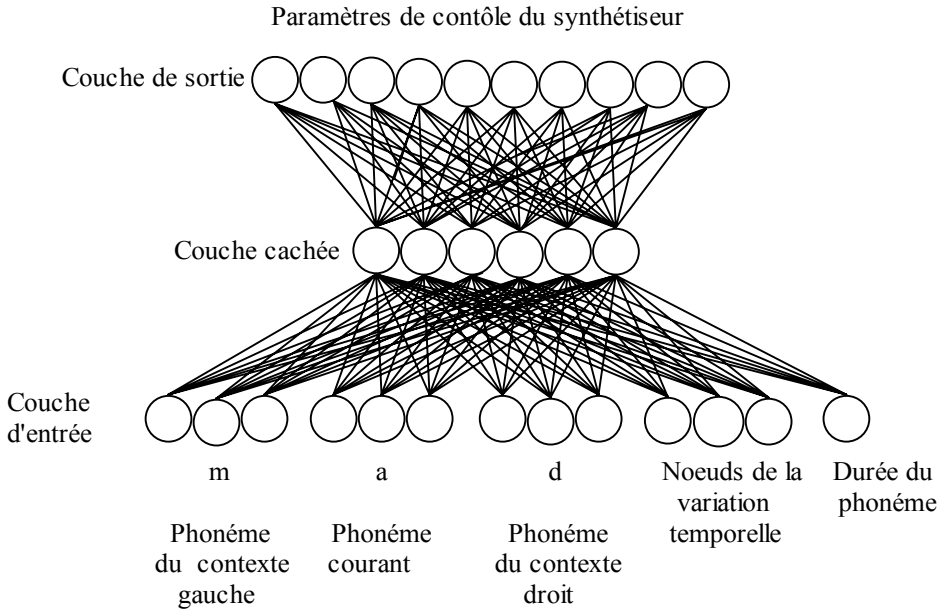


Figure 5.10 : Architecture du réseau de neurones phonétique-acoustique

5.5.2. Apprentissage et test

L'algorithme de rétropropagation du gradient est utilisé pour l'apprentissage du réseau de neurones phonétique-acoustique. Un bloc diagramme de la phase d'apprentissage est montré sur la figure 5.11.

L'un des problèmes qui apparaît durant l'apprentissage d'un réseau de neurones est le phénomène de sur-apprentissage. L'erreur d'apprentissage, tend à devenir très petite, mais quand de nouvelles données sont présentées au réseau, l'erreur devient importante. Le réseau a mémorisé les exemples d'apprentissages, mais n'a pas appris à faire la généralisation à de nouvelles situations.

Pour remédier à ce problème, une méthode qui permet d'améliorer la généralisation du réseau neuronal (méthode Early Stopping) a été utilisée. Nous avons divisé les données du corpus en trois parties : 60% pour l'apprentissage, 25% pour la validation et 15% pour le test. L'erreur de validation est surveillée durant le processus d'apprentissage. Tout comme l'erreur d'apprentissage, l'erreur de validation doit diminuer normalement durant la phase initiale de l'apprentissage, mais commence à augmenter lors du phénomène de sur-

apprentissage; à ce moment, l'apprentissage s'arrête, et les poids synaptiques et les biais correspondant au minimum de l'erreur de validation sont retenus. L'erreur sur l'ensemble de test n'est pas utilisée durant le processus d'apprentissage, mais peut être utilisée pour comparer les différents modèles. Il est également utile de tracer l'erreur de test durant le processus d'apprentissage.

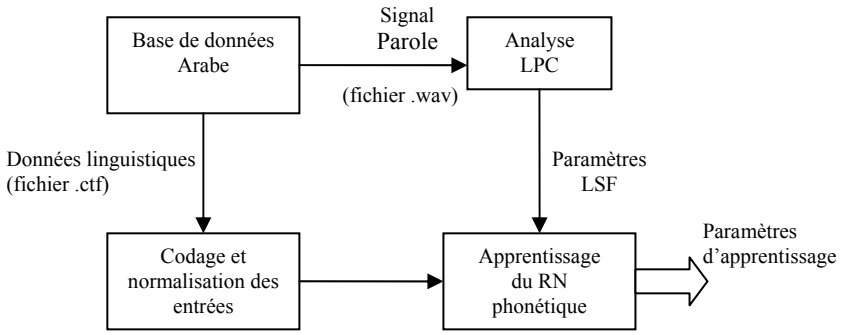


Figure 5.11 : L'étape d'apprentissage du réseau de neurones phonétique-acoustique

5.5.3. Lissage linéaire

En plus des propriétés d'interpolation des réseaux de neurones, nous avons implémenté un lissage linéaire des paramètres LSF pour produire des transitions lisses entre les phonèmes voisins. Nous appliquons une interpolation linéaire sur les coefficients LSF de quelques fenêtres au point de concaténation entre phonèmes adjacents [27, 148].

Soit L et R les segments gauche et droit au point de concaténation et X un vecteur LSF. Nous désignons par M_L et M_R le nombre des fenêtres du côté gauche et du côté droit au point de concaténation respectivement. Les paramètres LSF après lissage ($\hat{X}$) sont :

$$\begin{aligned} \hat{X}_L^{-i} &= X_L^i + (X_R^0 - X_L^0)\frac{M_L - i}{2M_L} \qquad 0 \le i \le M_L - 1 \\ \hat{X}_R^j &= X_R^j + (X_L^0 - X_R^0)\frac{M_R - j}{2M_R} \qquad 0 \le j \le M_R - 1 \end{aligned} \tag{5.15}$$

Où X_L^0 et X_R^0 sont les fenêtres à la fin de L et au début de R, c'est-à-dire exactement au point de concaténation. La fonction du lissage linéaire est montrée sur la figure 5.12, où M_L et M_R sont respectivement 2 et 3.

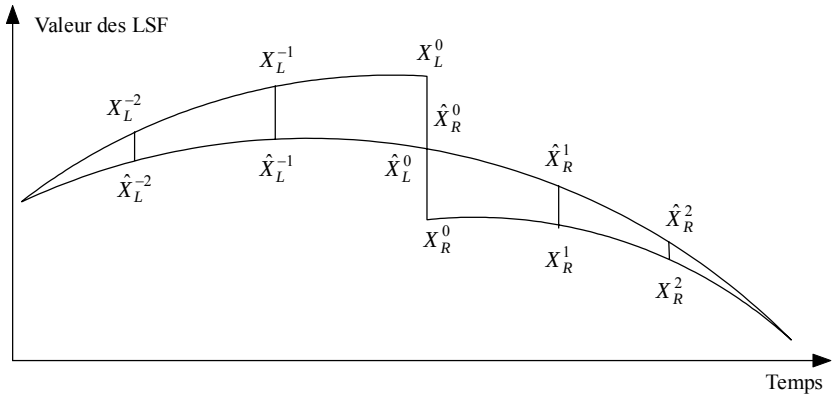


Figure 5.12 : Lissage linéaire des paramètres LSF au point de concaténation

5.6. Synthétiseur d'informations prosodiques à base de RN

La modélisation de la prosodie est une composante importante dans la synthèse de la parole à partir du texte. Dans notre système, trois réseaux de neurones standard à rétropropagation du gradient sont utilisés pour la génération de l'information prosodique. Ces réseaux de neurones sont décrits dans les sections 5.1 et 5.2. Pour leur apprentissage, validation et test, nous avons utilisé la base de données décrite dans le chapitre 4. Les contours intonatifs de chaque phrase de la base de données sont stylisés linéairement en utilisant l'algorithme MOMEL [149] (figure 5.13), et convertis de l'échelle linéaire Hz en échelle logarithmique semi-ton (st) en utilisant l'équation (5.16) [150] :

$$S_t = 12\left[\ln(Hz/100)/\ln(2)\right] \qquad (5.16)$$

qui atteint la valeur de référence $S_t = 0$ semi-tons à 100 Hz, $S_t = 12$ à 200 Hz et $S_t = -12$ à 50 Hz. Notons que la fonction coût utilisée pour l'apprentissage des différents réseaux de neurones est la l'erreur quadratique moyenne entre les valeurs de la sortie et la sortie désirée.

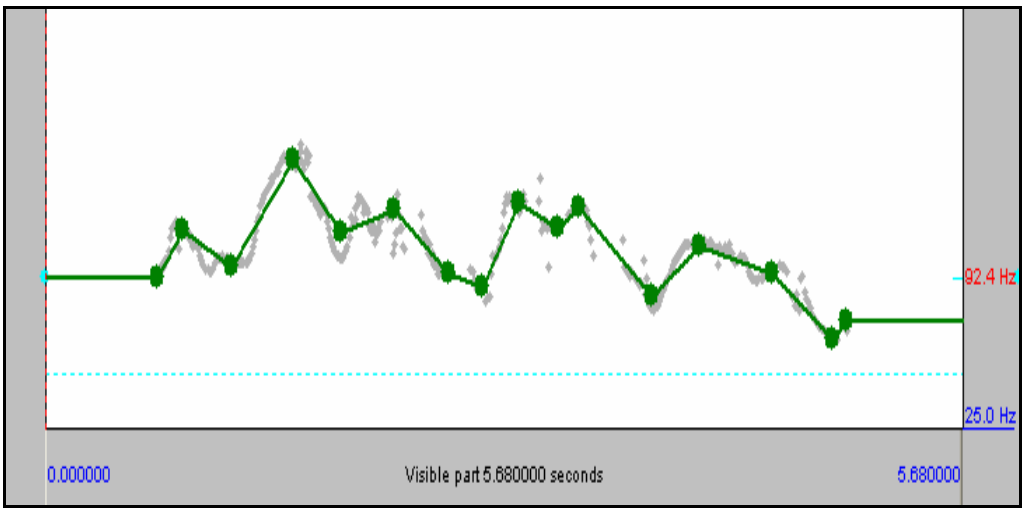


Figure 5.13 : La courbe intonative de la phrase « من رواد النهضة الحديثة في العالم العربي»
et sa version linéairement stylisée

5.6.1. Modélisation de la durée segmentale

La durée de chaque phonème de la phrase à synthétiser est générée en utilisant un réseau de neurones dont les paramètres d'entrées sont :

- l'identité du phonème courant qui est codée en utilisant le codage 1-parmi-n ;
- les phonèmes contextuels : nous avons considéré l'utilisation d'un phonème à gauche et un phonème à droite du phonème courant. Mais au lieu de coder leurs identités, nous avons considéré leurs classes phonétiques (voyelles : longue/courte, classes phonétiques des consonnes : occlusive, fricative, nasale, liquide, etc. En tenant compte du caractère voisé/non voisé du phonème) car il n'y a pas suffisamment d'exemples dans la base de données pour faire apprendre au réseau de neurones tous les contextes possibles. Un codage 1-parmi-n est utilisé pour coder les classes phonétiques des phonèmes contextuels ;
- le type des syllabes courante, gauche et droite, ainsi que leurs accents lexicaux. Ces paramètres sont codés en utilisant le codage 1-parmi-n ;
- la position du phonème dans la syllabe, la position de la syllabe dans le mot, et la position du mot dans la phrase. Ces paramètres d'entrée sont codés sous forme de nombres réels compris entre 0 et 1, avec 0 représentant la position initiale et 1 représentant la position finale ;
- le type du mot (mot grammatical ou lexical) qui est codé en utilisant un codage binaire.

Les 105 noeuds d'entrées du réseau de neurones reçoivent l'ensemble des caractéristiques d'entrée codées. La sortie est un seul neurone qui code linéairement la durée segmentale en valeurs comprises entre 0 et 1, et utilise la fonction d'activation sigmoïde. Ce réseau de neurones renferme une seule couche cachée avec 10 unités cachées de type tanhyp (tangente hyperbolique).

5.6.2. Modélisation du gain et de l'intonation

Pour générer les contours intonatifs et les niveaux d'énergie, nous avons choisi la syllabe comme unité de base. Pour un bon mappage, il est également important de fournir les informations contextuelles de la syllabe. Nous avons choisi une syllabe gauche (précédente) et une syllabe droite (suivante) de la séquence d'entrée. Les caractéristiques

d'entrée suivantes sont présentées au réseau de neurone au niveau de la syllabe pour chaque unité contextuelle :

1. le type de la syllabe codé en utilisant le codage 1-parmi-n ;
2. la durée de la syllabe normalisée entre 0 et 1 ;
3. l'information phonétique: la structure phonétique de la syllabe à traiter est codée comme suit :
    - la voyelle de la syllabe est présentée comme une entrée utilisant l'ensemble des phonèmes SAMPA Arabe et codée en utilisant le codage 1-parmi-n ;
    - les phonèmes voisins de la voyelle sont présentés en classes phonétiques (occlusive, fricative, nasale, liquide, semi-voyelle, etc.) en utilisant également le codage 1-parmi-n dans une fenêtre contextuelle symétrique d'un seul phonème des deux côtés.
4. l'information d'accentuation : nous avons utilisé le codage 1-parmi-n pour coder l'accent lexicale de la syllabe ;
5. l'information de la position : cette information est codée en utilisant un nombre réel compris entre 0 et 1. Cette information inclue la position :
    - de la voyelle dans la syllabe ;
    - de la syllabe dans le mot ;
    - du mot dans la phrase.

    Ainsi que le nombre de :
    - phonèmes dans une syllabe ;
    - de syllabes dans le mot ;
    - de mots dans la phrase.

Dans le cas du réseau de l'intonation, nous avons utilisé cinq valeurs analogiques de F_0 comme caractéristiques de sortie : F_0 au début de la syllabe, F_0 à la fin de la syllabe, et trois autres valeurs de F_0 uniformément réparties sur la durée de la syllabe (exprimée en dizaines de millisecondes).

Quelques paramètres importants dans l'architecture du réseau de l'intonation sont comme suit :

1. le nombre de neurones dans la couche d'entrée est égal à 216 ;
2. Une seule couche cachée avec 100 unités cachées et une fonction d'activation "tanhyp" ;
3. La couche de sortie a une fonction d'activation linéaire.

Les mêmes caractéristiques d'entrée décrites ci-dessus sont présentées à l'entrée du réseau neuronal du gain. Un paramètre représentant le niveau d'énergie est utilisé comme sortie (le maximum du logarithme de l'énergie), qui est normalisé entre 0 et 1. La couche cachée renferme 60 unités cachées avec une fonction d'activation "tanhyp" et la couche de sortie utilise une fonction d'activation "sigmoïde".

5.7. Excitation résiduelle

5.7.1. Dictionnaire des résiduels

Dans cette approche, notre objectif est de rendre plus naturelle la parole synthétique en utilisant une source d'excitation naturelle. Un modèle d'excitation résiduel est appliqué dans ce travail. Il est basé sur la modification et la concaténation de variantes d'excitation naturelles extraites à partir du signal vocal de la base de données en utilisant une méthode de filtrage LPC inverse, et stockées dans une bibliothèque de formes d'ondes. Une qualité de voix plus naturelle est obtenue si chaque phonème a plusieurs représentations dans le dictionnaire, chacune représentant un cas typique de ce phonème selon ses phonèmes de contexte gauche et droit. Dans ce cas les effets de la coarticulation sont inclus dans chaque représentation.

5.7.2. Modification et concaténation du résiduel

Les modifications prosodiques (du pitch et de la durée) sont faites dans la partie source du modèle source-filtre par la méthode TD-PSOLA [2, 67]. Il est plus avantageux d'appliquer les modifications TD-PSOLA au signal résiduel que de les appliquer sur le signal parole lui-même car le résiduel LPC est spectralement plat et il y a peu de corrélation dans une période de pitch. Ce signal d'erreur est dépourvu de la majorité des résonances du conduit vocal et par conséquent il évite les erreurs d'harmoniques de phase dans le cas de la modification et la concaténation. Après les modifications prosodiques des segments résiduels, ces derniers seront concaténés en utilisant une simple méthode de concaténation par addition-recouvrement.

5.8. Résultats expérimentaux

5.8.1. Performance du réseau de neurones phonétique-acoustique

La figure 5.14 montre l'évolution des erreurs d'apprentissage, de validation et de test durant le processus d'apprentissage du réseau phonétique-acoustique. L'apprentissage s'arrête après 245 itérations (à cause de l'utilisation de la méthode early-stopping), le taux

d'erreur obtenu étant inférieur à 2%. La figure 5.15 montre les coefficients LSF d'une phrase Arabe à partir des données de test. Ces coefficients obtenus à la sortie du réseau de neurones sont interpolés linéairement entre les phonèmes voisins en utilisant la technique de lissage décrite dans la section 4.3. Notons que la condition de stabilité (équation 5.15) du filtre de synthèse est vérifiée grâce à l'utilisation de l'expansion de largeur de bande et l'interpolation linéaire des coefficients LSF aux points de concaténation. Les trois premiers paramètres LSF les plus significatifs de la même phrase Arabe sont illustrés sur la figure 5.16. Nous remarquons, à partir de cette figure, que les paramètres LSF générés par le réseau neuronal suivent parfaitement bien ceux de la phrase originale, ce qui prouve l'efficacité de notre réseau de neurones. La figure 5.17 montre les ondes temporelles, les spectrogrammes et les contours F_0 de la phrase Arabe "مدينة بغداد". Elle présente la parole originale dans la partie (a), la parole synthétisée par le système dans la partie (b), la parole générée en utilisant une excitation résiduelle naturelle extraite à partir de la phrase originale dans la partie (c), et la parole générée en utilisant comme excitation un bruit blanc gaussien pour les sons non voisés et un train d'impulsions pour les sons voisés. Dans tous ces cas, l'information acoustique est générée en utilisant les durées des segments, l'énergie et l'intonation extraits de la phrase originale, illustrant seulement le comportement du réseau de neurones phonétique et celui du modèle d'excitation résiduelle.

La comparaison de la parole du cas (c) et la parole originale montre l'efficacité du réseau neuronal phonétique-acoustique car les spectrogrammes montrent des transitions lisses entre les phonèmes voisins. Les durées désirées des phonèmes et l'intonation sont également atteintes. Alors que la comparaison entre les cas (b) et (d) montre la robustesse du modèle d'excitation résiduelle.

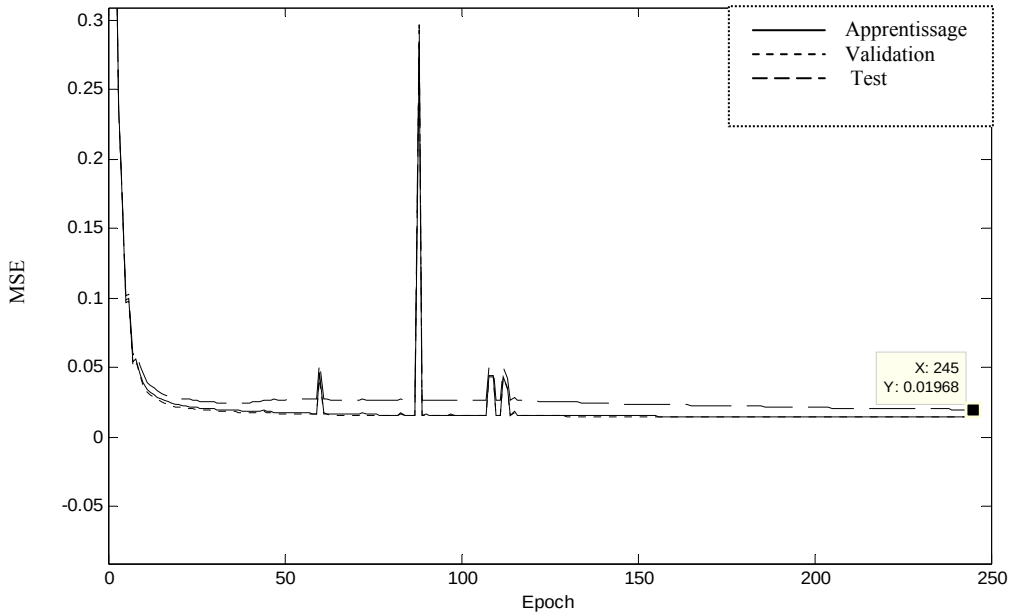


Figure 5.14 : Les performances du réseau de neurones phonétique durant l'apprentissage

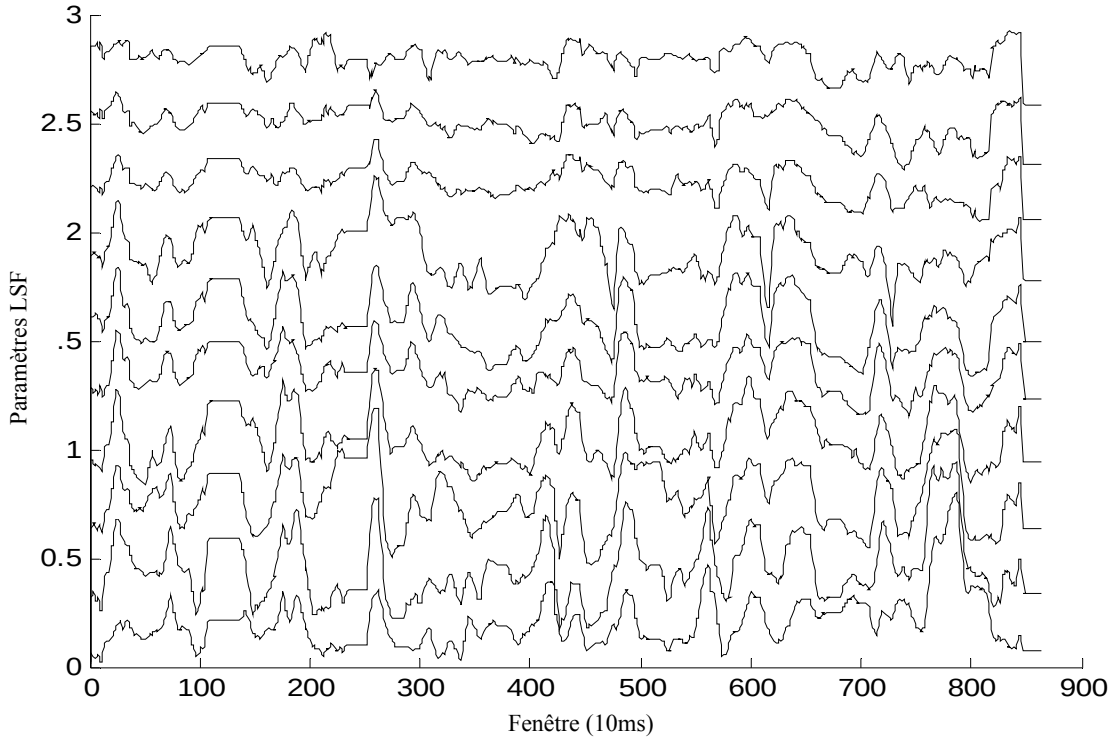


Figure 5.15 : Les paramètres LSF générés par le système pour une phrase Arabe test

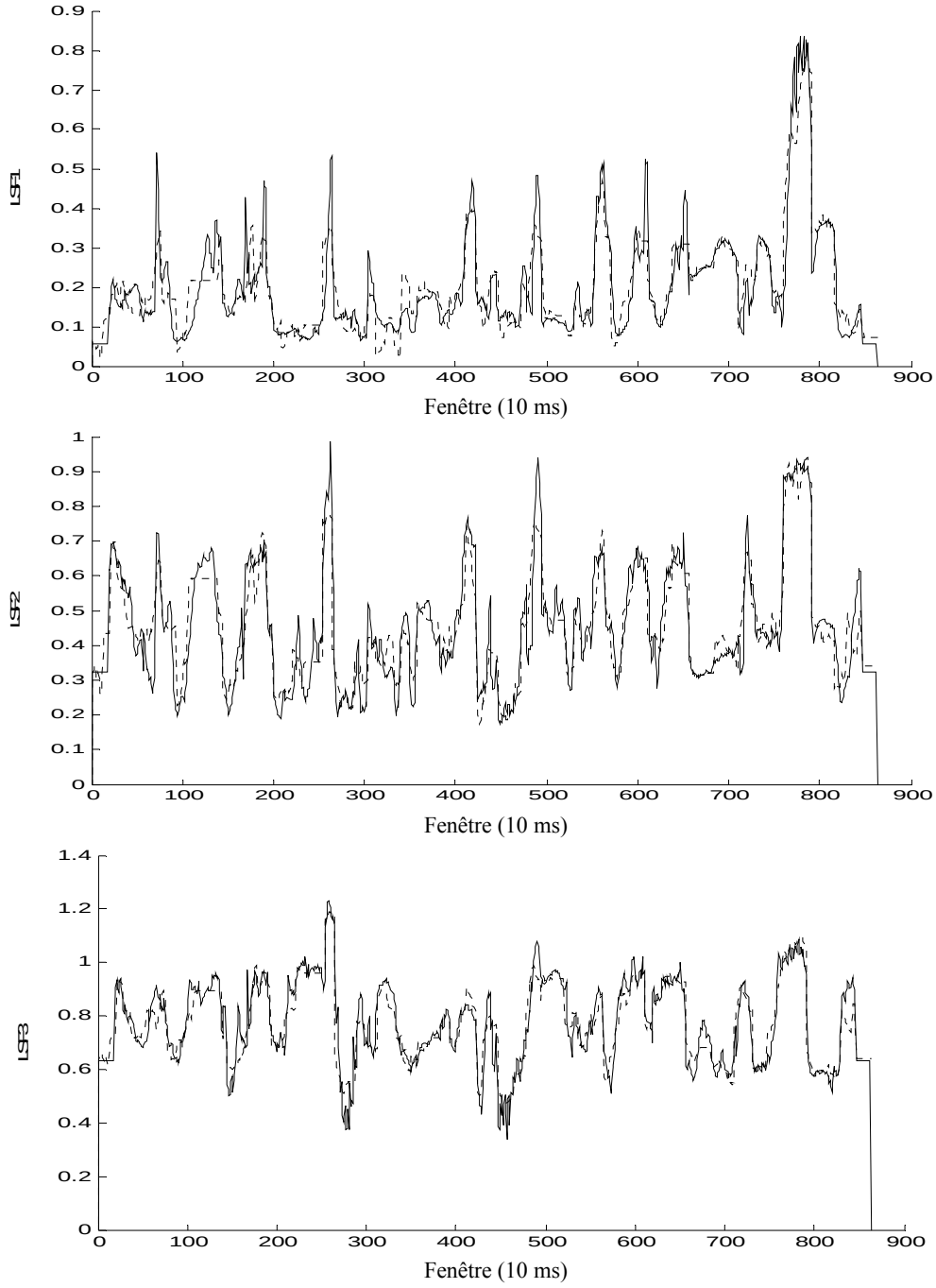


Figure 5.16 : Les trois premiers paramètres LSF de la phrase originale (continu) et la phrase synthétisée (discontinu)

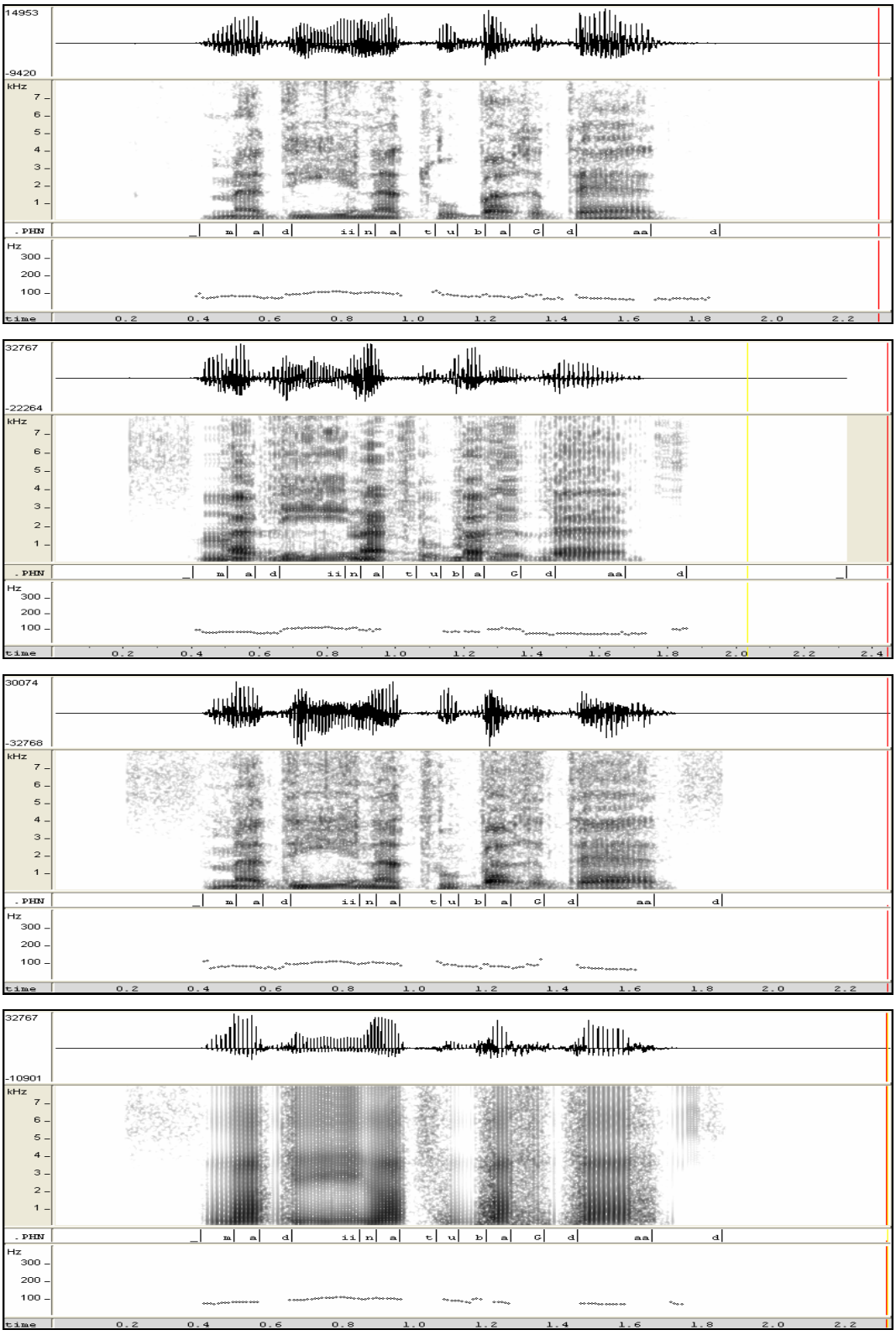


Figure 5.17 : De haut en bas nous avons : (a) parole originale, (b) parole synthétique utilisant le modèle d'excitation résiduelle proposé, (c) parole synthétique utilisant l'excitation résiduelle, (d) parole synthétique utilisant un simple train d'impulsions pour les sons voisés et un bruit blanc gaussien pour les sons non-voisés.

5.7.2. Performance des réseaux d'information prosodique

Dans le but d'évaluer objectivement la précision de prédiction entre les valeurs prédites par les modèles prosodiques proposés à base de réseaux de neurones et les valeurs réelles de chaque paramètre prosodique, la déviation standard (σ) et le coefficient de corrélation linéaire (r) sont calculés. Ces coefficients sont définis par les équations :

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \quad d_i = c_i - \overline{c}, \quad c_i = x_i - y_i, \quad \overline{c} = \frac{\sum_i c_i}{N} \tag{5.17}$$

et

$$r = \frac{V_{xy}}{\sigma_x \sigma_y} \tag{5.18}$$

Où σ_x et σ_y sont les coefficients de déviation standard de *x* et *y* respectivement, et

$$V_{xy} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{N}$$

Les valeurs moyennes de ces coefficients sont calculées dans le cas du paramètre prosodique F_0 car ce dernier renferme cinq valeurs de sortie. Le tableau 5.3 donne les résultats obtenus pour les trois paramètres prosodiques.

A partir de ce tableau, nous remarquons que l'évaluation objective de la durée segmentale donne une déviation standard de 17.5 ms pour les données de test. Une erreur de 6.25 % est mesurée pour le réseau neuronal de l'intonation; ceci est équivalent à 1.0496 semi-tons à 100 Hz et est bien au-dessous du seuil de parole 1.5 à 2 semi-tons [151]. Dans le cas de la synthèse du niveau d'énergie, une déviation standard de 5.14 dB est obtenue pour les données de test.

Tableau 5.3 : La précision de prédiction des trois paramètres prosodiques

	Données d'apprentissage		Données de test	
	r	σ	r	σ
Contours F_0	0.974	5.49 Hz	0.893	6.25 Hz
Niveau d'énergie	0.943	3.78 dB	0.879	5.14 dB
Durée	0.914	13.3 ms	0.831	17.5 ms

La figure 5.18 montre un segment typique d'un contour F_0 synthétisé. Il est clair à partir de cette figure que les trajectoires du pitch synthétique sont tout à fait proches de celles de la parole originale pour la majorité des syllabes. Les résultats de la synthèse du niveau d'énergie d'un exemple typique sont montrés sur la figure 5.19. Les trajectoires des niveaux d'énergie synthétiques sont également très proches de celles de la parole originale. Tous ces résultats prouvent que la génération des paramètres prosodiques par réseaux de neurones est une méthode simple mais très robuste.

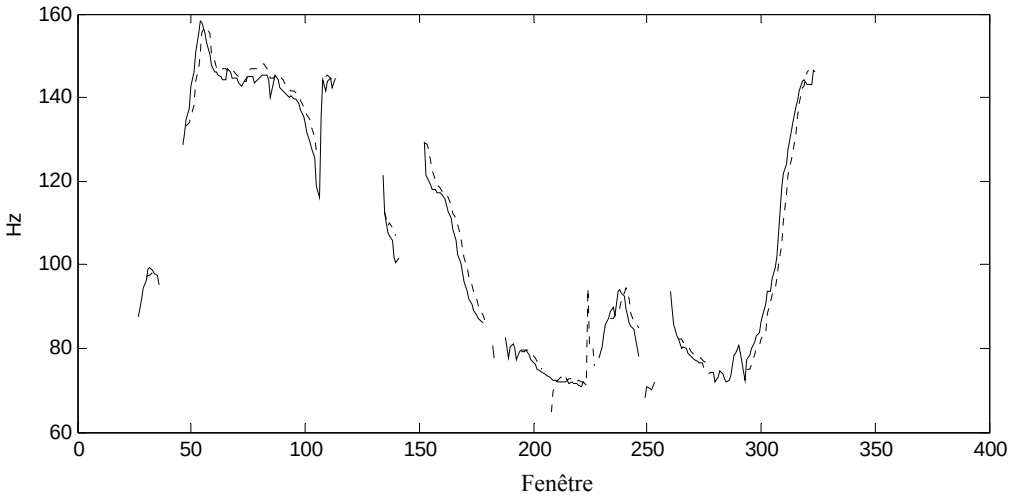


Figure 5.18. Les contours F_0 originaux (continu) et synthétisés (discontinu).

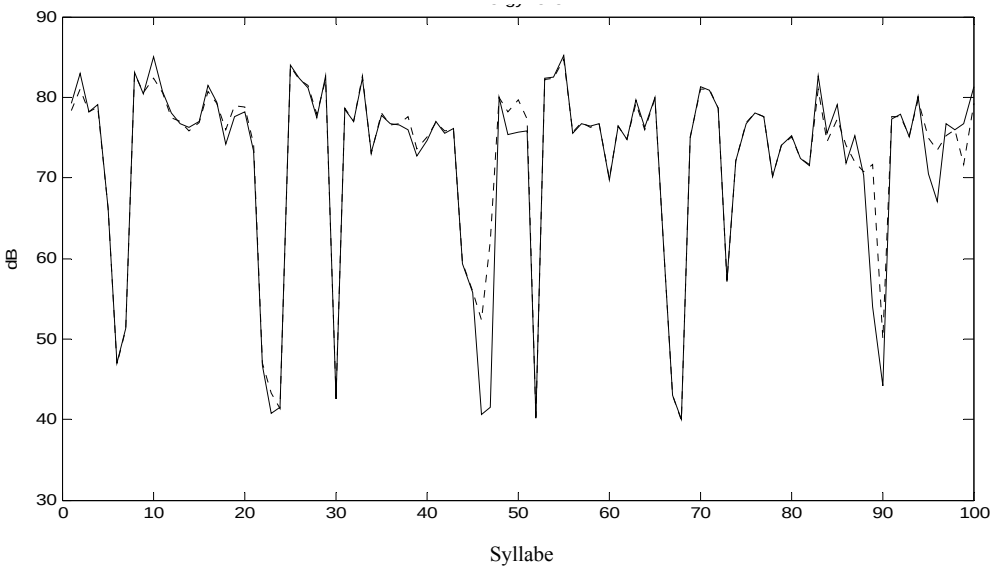


Figure 5.19 : Les contours des niveaux d'énergie originaux (continu) et synthétisés (discontinu).

5.7. Evaluation subjective et analyse statistique

L'évaluation de la parole synthétique est une étape importante dans le développement d'un système de synthèse vocale. Pour se faire nous avons réalisé un système de synthèse par sélection d'unités non uniformes afin de l'utiliser comme système de référence. Un test d'écoute formel a été effectué pour mesurer les performances de notre système à base de réseaux de neurones en comparaison avec le système de référence et les phrases originales de test. Cette section décrit le système de référence réalisé, les résultats du test d'écoute et l'analyse statistique de ces résultats pour savoir si les différences de moyenne obtenues pour les trois systèmes sont significatives ou elles sont dues au hasard des échantillons.

5.7.1. Synthétiseur par sélection d'unités pour Comparaison

Dans le but de mesurer les performances du système TTS Arabe à base de RN, un synthétiseur de parole basé sur une méthode de concaténation par sélection d'unités non uniformes (SUNU) est créé. Ce dernier sélectionne les segments à partir d'un corpus de parole en se basant sur deux formes de fonctions coût [7]. Le coût cible compare les segments sélectionnés et les segments désirés en utilisant certaines caractéristiques comme le pitch, la durée et l'énergie. Le coût de concaténation évalue la continuité spectrale perçue des segments potentiellement adjacents. Dans ce travail, nous avons utilisé les caractéristiques suivantes pour calculer le coût cible : la durée segmentale, la valeur moyenne de F_0 sur la longueur de l'unité, l'énergie moyenne, l'unité précédente, l'unité consécutive, et la position de l'unité dans la syllabe, le mot, et la phrase. Le coût cible est représenté sous forme d'une somme pondérée des différences entre les caractéristiques de l'unité cible et celles de l'unité candidat [7] comme c'est montré dans l'équation (5.19).

$$C^t(t_i, u_i) = \sum_{j=1}^{P} \omega_j^t C_j^t(t, u) \tag{5.19}$$

où p est le nombre des caractéristiques utilisées pour l'analyse du coût cible (dans notre cas p=8), ω_j^t est le poids associé à chaque caractéristique et C_j^t est la différence des caractéristiques entre une unité cible et une unité candidat de la base de donnés.

Pour calculer le coût de concaténation nous utilisons les vecteurs mel-cepstraux, la fréquence F_0 locale, et l'énergie locale. Les unités de la parole qui apparaissent consécutivement dans le corpus de la parole sont assignés un coût de concaténation nul ; la concaténation des unités consécutives de la base de données devrait nous fournir le point de concaténation le plus naturel et devrait donc être employée autant que possible. De

même, le coût de concaténation est représenté sous forme d'une somme pondérée des distances entre les caractéristiques d'une unité sélectionnée et l'unité qui la précède immédiatement comme c'est montré dans l'équation (5.20).

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{P} \omega_j^c C_j^c(u_{i-1}, u) \tag{5.20}$$

où *p* représente le nombre des paramètres utilisés pour l'analyse du coût de concaténation (dans notre cas p=3), ω_j^c est le poids associé à chaque paramètre, et C_j^c est la différence acoustique au point de concaténation entre deux unité de la parole.

Nous pouvons également employer cette mesure pour le couplage optimal des unités à concaténer [152]. Ceci signifie que nous pouvons également trouver le meilleur endroit pour joindre les deux unités. Au lieu de calculer le coût de concaténation sur les fenêtres de frontière au point de concaténation de deux segments voisins, nous changeons les frontières à l'intérieur d'une petite zone de quelques fenêtres (cinq fenêtres autour des frontières marquées des segments) afin de trouver le meilleur endroit pour la concaténation. Ce qui rend cette méthode peu sensible à de petites erreurs de segmentation. Les poids des coûts sont ajustés en utilisant une méthode d'apprentissage par régression linéaire proposée dans la référence [7], et un algorithme de Viterbi est utilisé pour trouver le chemin avec le coût minimal [7, 153].

Dix phrases originales sont retenues à partir de l'ensemble de test de notre base de données pour les besoins du test d'écoute ; les phrases restantes de la base de données sont employées comme inventaire d'unités acoustiques pour développer le système de synthèse par sélection d'unités. L'unité de base utilisée ici est le diphone, avec la possibilité d'utilisation de multiples phonèmes adjacents quand des assortiments appropriés sont trouvés dans la base de données [7, 154].

Après la sélection d'un segment de parole de la base de données, TD-PSOLA est utilisé pour ajuster le pitch et la durée de chaque phonème afin de les adapter aux propriétés désirées [2] (Les mêmes informations prosodiques générées par les réseaux de neurones décrits dans la Section 5.5 sont utilisées). Un simple algorithme basé sur une corrélation croisée est alors utilisé pour la concaténation des segments [155].

5.7.1. Test d'écoute formel et analyse statistique

Un test d'écoute a été effectué pour évaluer les performances de notre système de synthèse à base de RN en comparaison avec deux autres systèmes ; le premier est constitué des phrases test originales et le second correspond au système par sélection d'unités décrit ci-dessus. Les dix phrases de test retenues de la base de données ont été synthétisées par les deux synthétiseurs (synthétiseur à base de RN et synthétiseur par sélection d'unités). 19 adultes qui n'ont aucun problème d'audition et qui ont une bonne connaissance de l'Arabe ont participé dans ce test. Ils ont été invités pour évaluer chaque phrase synthétisée indépendamment pour l'intelligibilité, le naturel et la qualité globale de la voix. Pour chaque épreuve d'essai, une échelle d'évaluation de cinq niveaux MOS (Mean Opinion Score) a été présentée aux auditeurs pour choisir leurs jugements. Cette échelle va de un, pour la plus mauvaise qualité, jusqu'à cinq pour une qualité excellente. Le tableau 5.4 montre les notes moyennes d'opinion MOS qui ont été calculées pour chaque phrase par type de système. A partir de ce tableau nous remarquons que le système à base de RN est plus performant que le système SUNU, mais il n'est toujours pas aussi bon que le système original. Il est également intéressant de noter que les phrases test synthétisées par le système à base de RN sont jugées significativement plus naturelle que les phrases synthétisées par le système SUNU.

Tableau 5.4 : Résultats du test d'écoute formel

	Système RN	Système SUNU	Système original
Le naturel	4.8	4.25	4.95
L'intelligibilité	4.79	4.45	4.95
La qualité globale de la voix	3.7	3.375	5

Une analyse statistique a été également effectuée sur les résultats d'évaluation MOS obtenus pour les trois systèmes. Dans le cas de la qualité globale de la voix, le résultat d'ANOVA (ANalysis Of VAriance) à un facteur a indiqué qu'il y avait une différence significative entre les scores pour les trois systèmes ($F=93.67$; $df = 2,27$; $p<0.0001$). Les tests par la statistique t de Student ont indiqué que le système à base de RN est plus performant que le système SUNU en terme de la qualité globale de la voix ($t = 3.8806$; $df= 9$; $p = 0.0037$), mais il n'est pas aussi bon que le système original ($t = 11.7589$; $df = 9$; $p < 0.0001$). Pour l'évaluation du naturel des trois systèmes, le résultat de l'analyse d'ANOVA à un facteur indiquent qu'il y a une différence significative entre les notes moyennes des systèmes ($F=18.81$; $df=2,27$; $p<0.0001$). Les tests par la statique t montrent

que le système à base de RN génère une parole plus naturelle que le système SUNU (t=4.7143 ; df=9 ; p=0.0011). Cependant, la différence du naturel entre le système RN et le système original est considérée comme pas tout à fait statistiquement significative (t=1.964 ; df=9 ; p=0.0811), alors que la différence entre le système SUNU et le système original est extrêmement statistiquement significative (t = 5.7155 ; df = 9 ; P = 0.0003). Une analyse ANOVA additionnelle a été également effectuée pour l'intelligibilité des trois systèmes. Le résultat a indiqué que les trois systèmes différent de façon significative dans les préférences des auditeurs (F = 6.25 ; df = 2,27 ; p<0.0049). Les tests de la statistique t indiquent qu'il y a une préférence significative du système à base de RN par rapport au système SUNU (t = 3 ; df = 9; p = 0.015). De même, la préférence pour l'intelligibilité du système original par rapport au système SUNU est significative (t=2.849 ; df =9 ; p = 0.0191).

Cependant, la préférence du système original par rapport au système à base de RN en termes d'intelligibilité n'est statistiquement significative (t = 1 ; df = 9; p = 0.34).

5.8. Conclusion

Dans ce chapitre plusieurs composantes importantes d'un système TTS Arabe ont été décrites. Après avoir considéré la théorie de base des réseaux de neurones MLP, ainsi que leur méthode d'apprentissage, notre principal souci était l'utilisation de ces réseaux de neurones pour la génération des coefficients LSF d'un codeur LPC et les informations prosodiques afin de pouvoir produire une parole synthétique de haute qualité. Un modèle d'excitation résiduelle a été utilisé pour générer la source d'excitation du codeur LPC. Ce modèle consiste à extraire les segments résiduels requis à partir d'un dictionnaire, ensuite modifier leur prosodie avant de les concaténer. Les principaux problèmes rencontrés lors du développement de ce synthétiseur et les solutions proposées sont discutés en détail dans ce chapitre. Les résultats expérimentaux et les tests d'écoute formels montrent les performances du système TTS réalisé. En effet, notre système TTS à base de réseaux de neurones est jugé plus performant qu'un système TTS utilisant la concaténation par sélection d'unités.

CONCLUSIONS GENERALES ET PERSPECTIVES

L'objectif primaire du travail décrit dans cette thèse est de développer un système TTS Arabe de haute qualité. Une revue des principales méthodes de synthèse de la parole montrent que ces méthodes sont dépendantes de la langue étudiée et du locuteur. Elles nécessitent une étude énorme, beaucoup de travail et un espace mémoire très important pour le stockage des données, ce qui nous a incité à utiliser les réseaux de neurones artificiels dans notre système TTS. En effet, La synthèse de la parole par réseaux de neurones offre les avantages de portabilité de langue, de production d'une parole plus naturelle, d'un espace de stockage limité en comparaison avec les méthodes de synthèse classiques, aussi bien que sa capacité de produire des paramètres de synthèse raisonnables une fois confrontée à des situations non incluses dans les données d'apprentissage. Les systèmes TTS à base de réseaux de neurones peuvent aussi être configurés facilement pour d'autre voix en apprenant automatiquement à partir des bases de données existantes.

Pour la représentation du signal parole à la sortie de notre système de synthèse TTS, nous avons choisi la technique LP-PSOLA qui rassemble les avantages du modèle LPC et de l'algorithme TD-PSOLA. En effet, lors de la synthèse, un réseau neuronal génère les coefficients LSF d'un filtre de synthèse LPC. La source d'excitation de ce filtre est générée par un module qui consiste à extraire les segments résiduels nécessaires à partir d'un dictionnaire d'excitation, et ensuite modifier leur prosodie par l'algorithme TD-PSOLA avant de les concaténer.

La qualité d'un système de synthèse vocale à partir du texte dépend des techniques et des méthodes de synthèse, mais également du soin pris pour modéliser la prosodie, particulièrement en introduisant les effets intonatifs sur la parole synthétique. Quelques modèles existants pour la prédiction de la durée segmentale et des contours intonatifs ont été également examinés de près. Les problèmes généraux dans la modélisation de la prosodie résident dans la relation non linéaire et floue qui existe entre la représentation symbolique et discrète de la parole et sa manifestation réelle comme signal continu

variable. Fondamentalement, on doit développer une méthodologie pour associer un ensemble de représentations linguistiques, paralinguistiques et émotives aux paramètres prosodiques de la parole. La solution qui a été proposée consiste à utiliser les réseaux de neurones pour remplir cette tâche. Ainsi, trois réseaux de neurones ont été utilisés pour la modélisation des trois paramètres prosodiques principaux : la fréquence fondamentale F_0, la durée segmentale et l'énergie.

Afin de pouvoir atteindre nos objectifs, nous avons fait plusieurs expériences avec différentes topologies des réseaux de neurones pour déterminer leurs tailles optimales. Nous avons testé également différents types d'entrée pour déterminer les paramètres nécessaires pour un problème donné. Des solutions pratiques ont été proposées pour les problèmes rencontrés lors du développement de ce synthétiseur. Nous citons l'expansion de la largeur de bande et l'interpolation linéaire des coefficients LSF aux points de concaténation des phonèmes.

Cette thèse présente également les différentes étapes entreprises pour l'élaboration d'une base de données prosodiques de l'Arabe. Cette base de données a été utilisée dans l'apprentissage de tous les réseaux de neurones utilisés dans ce travail. Elle contient des informations très fiables pour couvrir l'ensemble d'apprentissage. Cependant, si quelques informations ne sont pas couvertes par cette base de données, nous pouvons surmonté ce problème en utilisant une méthode de codage appropriée présentant aux réseaux de neurones une structure d'entrée qui reflète des similitudes parmi les classes d'entrée.

Les résultats expérimentaux et les tests d'évaluation subjective montrent que le système TTS à base de réseaux de neurones est plus performant pour un ensemble d'apprentissage relativement petit, qu'un système TTS par sélection d'unités non uniformes utilisant le même corpus comme inventaire d'unités acoustiques. Il reste cependant, quelques améliorations qui pourraient faire l'objet d'effort et d'investissement. A notre point de vue, ces résultats peuvent encore être améliorées à travers l'étude et le traitement des points suivants :

- la base de données doit être élargie pour inclure plus de variations d'intonations que celle présente dans la version courante, aussi bien que pour inclure autant de contextes phonétiques et autant d'environnements acoustiques que possible ;

- des améliorations dans le modèle d'excitation résiduelle sont aussi possibles. En particulier, un peu plus de robustesse est nécessaire dans la concaténation et les algorithmes de lissage appliqués aux segments résiduels ;
- l'utilisation d'un autre modèle de parole tel que le modèle HNM qui offre la possibilité de lissage spectrale aux points de concaténation des segments, une bonne adaptation de la prosodie requise, ainsi que l'avantage de ne plus avoir besoin d'un dictionnaire d'excitation.

ANNEXE A
ABRÉVIATIONS

ANOVA	: ANalysis Of Variance
AR	: Auto-Regressif
AS	: Arabe Standard
BD	: Base de données
CELP	: Codebook Excited Linear Prediction
CGI	: Glottal Closure Instants
DTW	: Dynamic Time Wrapping
FD-PSOLA	: Frequency Domain PSOLA
HMM	: Hidden Markov Model
HNM	: Harmonic Plus Noise
LAR	: Log Area Ratio
LP	: Linear Prediction
LPC	: Linear Predictive Coding
LP-PSOLA	: Linear-Predictive PSOLA
LSF	: Line Spectral Frequencies
MBE	: Multi-Band-Excited
MBROLA	: Multi-Band Resynthesis pitch synchronous OLA
MLP	: Multi Layer Perceptron
MLPC	: Multipulse Linear Prediction Coding
MOS	: Mean Opinion Score
PMC	: Perceptron MultiCouche
PSOLA	: Pitch-Synchronous OverLap and Add
RAP	: Reconnaissance Automatique de la Parole
RBF	: Radial Basis Function
RELP	: Residual Excited Linear Prediction
RN	: Réseau de Neurones
SUNU	: Sélection d’Unités Non Uniformes

SVF : Spectral Variation Function

TAP : Traitement Automatique de la Parole

TD-PSOLA : Time-Domain PSOLA

ToBI : Tones and Break Indices system

TOP : Transcription Orthographique-Phonétique

TTS : Text-To-Speech

WLP : Warped Linear Prediction

ANNEXE B
THEORIE DE LA PREDICTION LINEAIRE

Les études qui ont été faites sur le mécanisme de la phonation, montrent que le signal vocal *x(n)* dans une fenêtre ($1 \leq n \leq N$) est produit par un système dont la transmittance est approximativement de la forme g/A(z) ; ce système serait soumis à une excitation u(n) sous forme d'un train périodique d'impulsions d'amplitude unité pour les sons voisés ; et pour les sons non voisés l'excitation est un bruit blanc (figure 1.8).

La transmittance g/A(z) est celle d'un filtre polynomial (ou tous pôles). On pourrait écrire alors:

$$X(z) = g.\frac{U(z)}{A(z)} \quad \Leftrightarrow \quad X(z).A(z) = g.U(z) \tag{B.1}$$

Où le polynôme A(z) est de la forme :

$$A(z) = \sum_{i=0}^{p} a_i z^{-i}$$

Avec $a_i(0)=1$, ou encore

$$A(z) = 1 + \sum_{i=0}^{p} a_i z^{-i} \tag{B.2}$$

Dans le domaine temporel, on aura l'expression correspondante :

$$x(n) + \sum_{i=1}^{p} a_i .x(n-i) = g.u(n) \tag{B.3}$$

qui exprime qu'un échantillon quelconque *x(n)* est une combinaison linéaire des *p* échantillons qui le précèdent. Ce modèle de production d'un signal est appelé « autorégressif » ou tout simplement modèle AR.

- les coefficients a_i son appelés coefficients de prédiction linéaire ;
- le coefficient g est appelé gain du système ;

- p est appelé ordre de prédiction.

Estimation du modèle

On peut définir une prédiction ou estimation de chaque échantillon $x(n)$ à partir des p échantillons qui le précèdent :

$$\hat{x}(n) = \sum_{i=1}^{p} \hat{a}_i . x(n - i) \tag{B.4}$$

Où les $\hat{a}_i$ (i=1,…,p) sont les estimés des coefficients de prédiction a_i .

L'erreur commise par la prédiction vaut :

$$e(n) = x(n) - \hat{x}(n) \tag{B.5}$$

$$e(n) = x(n) - \sum_{i=1}^{p} \hat{a}_i . x(n - i)$$

$$e(n) = \sum_{i=0}^{p} \hat{a}_i . x(n - i)$$

avec $\hat{a}_0 = 1$.

On remarque que si $\hat{a}_i = a_i$ l'erreur coïncide avec l'excitation à un facteur près :

$$e(n) = g.u(n) \tag{B.6}$$

On abandonnera désormais l'indice (^) pour désigner les coefficients estimés.

Estimer le modèle AR revient à estimer les coefficients a_i , c'est-à-dire trouver les coefficients optimaux.

On a :

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^{p} a_i . x(n - i) \tag{B.7}$$

On définit l'énergie résiduelle de prédiction par la somme :

$$E = \sum_{n=-\infty}^{+\infty} e^2(n) \tag{B.8}$$

Le critère usuel pour l'optimisation est la minimisation de l'énergie résiduelle de prédiction. L'énergie résiduelle peut s'écrire sous la forme :

$$E = \sum_{n} [x(n) - \sum_{i=1}^{p} a_i . x(n-i)]^2 \tag{B.9}$$

Trouver les a_i qui minimisent E revient à annuler les dérivées partielles de E par rapport à ces coefficients.

$$\frac{\partial E}{\partial a_j} = -2 \sum_{n=-\infty}^{+\infty} \left[x(n) - \sum_{i=1}^{p} a_i . x(n-i) \right] . x(n-j)$$

$$= -2 \sum_{n=-\infty}^{+\infty} x(n-j) . \left[x(n) - \sum_{i=1}^{p} a_i . x(n-i) \right] \quad \text{pour j} = 1,...,\text{p}$$

ou encore,

$$\frac{\partial E}{\partial a_j} = 2 \left\{ \sum_{i=1}^{p} \left[a_i . \sum_{n} x(n-i).x(n-j) \right] - \sum_{n} x(n-i).x(n-j) \right\} \tag{B.10}$$

En posant m=(n-i), $\frac{\partial E}{\partial a_j}$ devient :

$$\frac{\partial E}{\partial a_j} = 2 \left\{ \sum_{i=1}^{p} a_i . \sum_{m} x(m).x(m+i-j) - \sum_{n} x(n).x(n-j) \right\} \tag{B.11}$$

$$\frac{\partial E}{\partial a_j} = 0 \Rightarrow \sum_{i=1}^{p} a_i . \sum_{m} x(m).x(m+i-j) - \sum_{n} x(n).x(n-j) \quad pour \ j = 1,...,p$$

Ce sont les équations de YULE-WALKER. On sait que pour un signal stationnaire, la fonction d'autocorrélation vérifie :

$$R(k) = R(-k) = \sum_{l} x(l).x(l+k) \tag{B.12}$$

Donc les équations de YULE-WALKER deviennent :

$$\sum_{i=1}^{p} a_i R(i - j) = R(j) \tag{B.13}$$

En utilisant la notation matricielle, on aura :

$$\begin{bmatrix} R(0) & R(1) & \cdots & \cdots & R(p-1) \\ R(1) & R(0) & & & R(p-2) \\ \vdots & & \ddots & & \vdots \\ R(p-2) & & & \ddots & R(1) \\ R(p-1) & \cdots & \cdots & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ a(p-1) \\ a(p) \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p-1) \\ R(p) \end{bmatrix}$$

Donc :

$$R^{(p,p)}.A = R^{(p)} \tag{B.14}$$

La matrice $R^{(p,p)}$ est dite TOEPLITZ puisqu'elle vérifie :

- la symétrie ;
- elle est définie positive ;
- les éléments situés sur une diagonale parallèle à la diagonale principale sont identiques.

Plusieurs algorithmes ont été proposés pour la résolution de ce système. Le plus utilisé est celui de Levison-Durbin qui permet de résoudre le système par une récursion sur l'ordre de prédiction P.

REFERENCES

1. Dutoit, T., “An Introduction to Text-to-speech Synthesis for the French Language”, Thèse de Doctorat. Faculté Polytechnique de Mons, Belgique, (1997).

2. Moulines, E. and Charpentier, F., “Pitch synchronous waveform processing techniques for a Text-To-Speech synthesis using diphones”, Speech Communication, 9(5,6), (1990), pp. 453-467.

3. Dutoit, T., “Introduction au traitement numérique de la parole”, Faculté Polytechnique de Mons, Belgique, 1ère édition, (2000).

4. Holmes, J.N. et al., “Speech Synthesis by Rule, language and speech”, 7, (1964) pp.127-143.

5. Dixon, N. R., Maxey, H. D., “Terminal analog synthesis of continuous speech using the diphone method of segment assembly”, IEEE Trans. On Audio and Electroacoustics, AU-16, (1976), pp.40-50.

6. Guerti, M., “Contribution à la synthèse de la parole en Arabe Standard”, XVIèmes Journées d’Etudes sur la Parole (JEP), Société Française d’Acoustique, Hammamet, Tunisie, (5-9 Octobre 1987), pp. 290-293.

7. Hunt, A. J. and Black, A. W., “Unit selection in a concatenative speech synthesis system using a large speech database”, in Proc. ICASSP’96, (1996), pp. 373–376.

8. Boite, R., Boulard, H., Dutoit, T., Hancq, J. & Leich, H., “Traitement de la parole, chapitre : synthèse de la parole à partir d’un texte”, Collection Electricité, Presses polytechniques et universitaires romandes, (2000), pp. 345-441.

9. Guerti, M., “Speech synthesis by rule”, 8th International Conference on Computer Theory and Applications ICCTA’98, IEEE (Alexandra Chapter), Alexandria - EGYPT, III.12-III.15, (15-17 September 1998).

10. Baloul, S., “Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé”, Thèse de doctorat, université du Maine, Le Mans, France, (2003).

11. Tuerk, C., and Robinson, T., “Speech Synthesis using neural networks trained on cepstral coefficients”, Proc. Eurospeech'93, Berlin, (September 1993), pp. 1713-1716.

12. Weijters, T., and Thole, J., “Speech synthesis with artificial neural networks”, Proc. ICNN'93, San Francisco, (March 1993), pp.1764-1769.

13. Karaali, O., Corrigan, G., Gerson, I., and Massey, N., “Text-to-speech conversion with neural networks: A recurrent TDNN approach”, In Proc. Eurospeech'97, Rhodes, Greece, (1997), pp.561-564.

14. Xiang, Z., BI, G., “A neural network model for Chinese speech synthesis”, in Proc. IEEE international Symposium on circuits and Systems, Vol. 3, (1990), pp. 1859-1862.

15. Cawley, G.C., “The application of neural networks to phonetic modelling”, Phd thesis, University of Essex, (March 1996).

16. Chouireb, F. and Guerti, M., “Towards a high quality Arabic speech synthesis system based on neural networks and residual excited vocal tract model”, Signal, Image and Video Processing Journal, DOI : 10.1007/s11760-007-0038-z, ISSN : 1863-1703 (Print) 1863-1711 (Online), Springer London, (October 2007).

17. Tao, J., Cai, L. and Tropf, H., “An optimised neural network based prosody model of Chinese speech synthesis system”, Proc. of IEEE TENCON'02, (2002), pp. 477-480.

18. Farrokhi, A. and Ghammaghami, S., “Predication of prosodic data in Persian text-to-speech systems using recurrent neural network”, Electronics letters IEE, vol. 39, No. 25, (December 2003).

19. Teixeira, J. P. and Freitas, D., “Segmental durations predicted with a neural network”, Eurospeech -Geneva, 2003, pp.169-172.

20. Vainio, M., “Artificial neural network based prosody models for Finnish Text-to-speech synthesis”, University of Helsinki, Department of Phonetics, Finland, (2001).

21. Chen, S. H., Hwang, S. H. and Wang, Y. R., “An RNN-based prosodic information synthesizer for Chinese text-to-speech”, IEEE Trans. Speech Audio Processing, vol.6, (May 1998), pp. 226-239.

22. Erdem, C. and Zimmermman, H. G., “A data-driven method for input feature selection within neural prosody generation”, Proc. ICASSP, Vol. 1, (2002), pp. 477-480.

23. El-Ani, M., “Arabic Phonology: An acoustical and physiological investigation” , Mouton, The Hague, Paris, (1970).

24. Guerti, M., “Contribution à la synthèse de la parole Arabe Standard (Synthèse par dihpones et technique de prédiction linéaire)”, Thèse de Magister en Electronique Acoustique et physiologique de la parole, ILP. Alger et CENT Lannion, (Mars 1984).

25. El-Imam, Y.A., “Unrestricted vocabulary Arabic speech synthesis system”, IEEE Trans. Acoustic, Speech signal Processing, ASSP-37 (12), (1998), pp.1829-1845.

26. En-najjary, T., “Conversion de voix pour la synthèse de la parole”, Thèse de Doctorat de l’université de Renne 1, (Mars 2005).

27. DuToit, T., “An Introduction to Text–To–Speech Synthesis”. Kluwer, (1996).

28. Klatt, D. H., “Review of text–to–speech conversion for English”, Journal of the Acoustical Society of America, vol. 82, (September 1987), pp. 737 – 793.

29. Lucassen, J., and Mercer, R., “An information theoretic approach to the automatic determination of phonemic base forms”, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, (1984), pp.42.5.1-42.5.4.

30. Sejnowski, T., and Rosenberg, C., “NetTalk : A parallel network that learns to read aloud”, Technical report JHU/EECS-86/01, Johns Hopkins University, MD, (1986).

31. Zemirli, Z., “SYNTHAR+ : Synthèse vocale sous MULTIVOX”, Techniques et Science informatique, 17(6), (1998).

32. Saroh, A., Brusset, J. et Tihoni, J., “Vers une production automatique de textes phonétiques pour l’arabe standard à partir de sa graphie”, Actes des 18èmes Journées d’études sur la parole, Montréal, (1990), pp.305-309.

33. Hendessi, F., Ghayoori, A., Aaron, T., Gulliver: “A speech synthesizer for Persian text using a neural network with a smooth ergodic HMM”. ACM Trans. Asian Lang. Inf. Process. 4(1): 38-52, (2005).

34. Ishizaka, K. and Flanagan, J. L., “Synthesis of voiced sounds from a two-mass model of the vocal chords”, Bell systems Technology Journal, 50:1233-1268, (1972).

35. Flanagan, J. L. and Ishizaka, K., “Computer model to characterise the air volume displaced by the vibrating vocal chords”, Journal of the Acoustical Society of America, 63:1559-1565, (1978).

36. Maeda, S., “Improved articulatory model”, Journal of the Acoustical Society of America, vol. 84, no. S1, p. S146, (1988).

37. Perrier, P., Boë, L. J. & Sock, R., “Vocal-tract area function estimation from midsagittal dimensions with CT scans and a vocal-tract cast”, J. Speech Hearing Research 35, 53-67, (1992).

38. Stevens, K. N., Kasowski, S. and Fant, G., “An electrical analog o the vocal tract”, Journal of the Acoustical Society of America, 25:734-42, (1953).

39. Rye, J. M. and Holmes, J. N., “A versatile software parallel-formant speech synthesizer”, JSRU research report 1016, Joint Speech Research Unit, Cheltenham, UK, (1982).

40. Cawley, G., “The Application of Neural Networks to Phonetic Modelling”. PhD. Thesis, University of Essex, England, (1996).

41. Fant, C. G. M., “On the predictability of formant levels and spectum envelopes from formant frequencies”, for Roman Jackobson, (1956), pages 109-120.

42. Klatt, D. H. “Software for a cascade/parallel formant synthesiser”, Journal of the Acoustical Society of America, vol. 67, (1980, pp. 971 – 995.

43. Allen, J., Hunnicutt, S., Klatt D., “From Text to Speech: The MITalk System”. Cambridge University Press, Inc, (1987).

44. Magnusson, L., Blomberg, M., Carlson, R., Elenius, K. and Granstrom, B., “Swedish Speech Researchers Team-up With Electronic Venture capitals”, Speech Technology, 2, pp. 15-24, (1984).

45. Groner, G.F., Bernstein, J., Ingberg, E., Rearlman, J. and Toal, T. “A real-time text-to-speech converter”, speech Technology, 1, PP. 73-76, (1982).

46. Klatt, D., “The Klattalk Text-to-Speech Conversion System”, Proceedings of ICASSP 82 (3): 1589-1592, (1982).

47. Bruckert, E., Minow, M. and Tetschner, W., “Three-Tiered software and VLSI aid developmental system to read text aloud”, Electronics, 56, pp. 133-138, (1983).

48. Conkie, A., “Robust unit selection system for speech synthesis”, Acoustical Society of America meeting, Berlin, (1999).

49. Coorman, G., Fackrell, J., Rutten, P. & Van Coile, B. “Segment selection in the L&H Real Speak laboratory TT system”, ICSLP, Bejiing, (2000).

50. Rutten, P., Coorman, G., Fackrell, J. & Van Coile, B., “Issues in corpus based speech synthesis”, Seminar IEE, State of the art in speech synthesis, Savoy Place, (2000), pp. 16/1-16/7.

51. Pierre-Yves Le Meur, “Synthèse de la parole par unités de taille variable”, Thèse de docteur, ENST (Télécom Paris), (1996).

52. Stylianou,Y., “Modèles hamonique plus bruit combinés avec des méthodes statistiques pour la transformation de la parole et du locuteur”, thèse de doctorat, ENST (Télécom Paris) , (1996).

53. Dutoit, T., Pagel, V., Pierret, N., Van Der Vreken, O., Bataille, F., “The Mbrola Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes”, Proc. ICSLP 96, Philadelphia, (1996).

54. Page, J. and Breen, A., “The laureate text-to-speech system, architecture and applications”, BT Technical Journal, vol. 14, (January 1996), pp. 57-67.

55. Beutnagel, M., Conkie, A., Schroeter, J. and Syrdal, A., “The at&t next-gen TTS system”, in Joint Meeting of ASA, EAA, and DAGA, (Berlin, Germany), (March 1999).

56. Donovan, R., “Trainable Speech Synthesis”, PhD. Thesis. Cambridge University Engineering Department, England, (1996).

57. Emerard, F., “Synthèse par diphones et traitement de la prosodie”, thèse de Doctorat, Université de Grenoble III, (1997).

58. Courbon, J.L. and Emerard, F., “SPARTE: A text-to-speech machine using synthesis by diphones”, Proc. ICASSP’82, Paris, (1982), pp. 1597-1600.

59. Chenfour, N., “Réalisation d’un système de synthèse de la parole arabe à partir du texte par concaténation des di-syllabes”, Thèse de Doctorat 3ème cycle, Université Mohamed V, Rabat (Maroc), (Juillet 1997).

60. Islam, T., “Interpolation of linear prediction coefficients for speech coding”, Master thesis, M. Gill University, Montreal, Canada, (April 2000).

61. Kabal, P. & Ramachandran, R.P., “The computation of line spectral frequencies using Chebyshev polynomials”, IEEE trans. Acoustics, Speech, Signal Processing, Vol. ASSP-34, (Dec. 1986), pp. 1419-1426.

62. Laine, U., Karjalainen, M., Altosaar, T., “Warped Linear Prediction (WLP) in Speech Synthesis and Audio Processing”, Proceedings of ICASSP94 (3): 349-352, (1994).

63. Karjalainen, M., Altosaar, T. Vainio, M., “Speech Synthesis Using Warped Linear Prediction and Neural Networks”, Proceedings of ICASSP98, (1998).

64. Atal, B.S. and Remde, J.R., “A new model of LPC excitation for producing natural-sounding speech at low bit rates”, Proc. ICASSP’82, Paris, (1982), pp. 614-617.

65. Campos, G., Gouvea, E., “Speech Synthesis Using the CELP Algorithm”, Proceedings of ICSLP 96 (3), (1996).

66. Charpentier, F.J. and Stella, M.G., “Diphone synthesis using an overlap-add technique for speech waveforms concatenation”, Proc. ICASSP’86, Tokyo, (1986), pp. 2015-2018,.

67. Chouireb, F. & Guerti, M., “Etude et Application des techniques LPC et TD PSOLA pour l’analyse/ modification/synthèse de la parole”, International Conference on Electrical and Electronics Engineering ICEEE’2004Université Amar Telidji Laghouat (Algérie), Special Issue ISSN 1112-4652, (April, 24-26 2004), pp. 244-250.

68. Chouireb, F. and Guerti, M., “Amélioration de la qualité de la parole synthétique en Arabe Standard”, Algerian Journal Of Technology, Série B, 16(1)(2006), ISSN 1111-357x.

69. Charpentier, F., Moulines**,** E., **“**Text-to-speech algorithms based on FFT synthesis”, in International Conference on Acoustics, Speech, and Signal Processing, ICASSP-88, vol.1, pp. 667-670, (11-14 Apr 1988).

70. DuToit, T. and Leich, H., “MBR–PSOLA: Text–to–speech synthesis based on an MBE re–synthesis of the segments database”, Speech Communication, vol. 13, pp. 435 – 440, (1993).

71. McAulay, R. & Quatieri, T., “Speech coding and synthesis, Sinusoidal coding”, Elsevier science, 1995, p. 121-173.

72. Stylianou, Y., “Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modifications”, Thèse, Telcom Paris, (Janvier 1996).

73. Syrdal, A., Stylianou, Y., Garisson, L., Conkie, A. and Schroeter, J., “TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis”, Proc. IEEE Int. Conf. Acoust., Speech, signal Processing, (1998), pp. 273-276.

74. Fries, G., “Hybrid time- and frequency-domain speech synthesis with extended glottal source generation”, Acoustics, Speech, and Signal Processing. ICASSP-94, Page(s):I/581 - I/584 vol.1, (19-22 Apr 1994).

75. Donovan R. and Woodland, P., “Automatic speech synthesizer parameter estimation using HMMs”, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, (May 1995), pp. 640-643.

76. Huang, X. Acero, A., Adcock, J., Hon, H., Goldsmith, J., Liu, J. and Plumpe, M., “Whistler: A trainable text-to-speech system”, in Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA, vol.4, pp. 2387-2390, (Oct. 1996).

77. Renzepopoulos, P., Kokkinakis, G., “Multilingual Phoneme to Grapheme Conversion System Based on HMM”, Proceedings of ICSLP 92 (2): 1191-1194, (1992).

78. Takuda, K., Kobayashi, T. and Imai, S., “Speech parameter generation from HMM using dynamic features”, Proc. ICASSP’95, Detroit, 1995, pp.660-663.

79. Takuda, K., Masuko, T., Yamada, T., Kobayashi, T. and Imai, S. “An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features”, Proc. Eurospeech’95, Madrid, (1995), pp. 757-760.

80. Sluijter, A., Terken, J.M.B., “Beyond sentence prosody: Paragraph intonation in Dutch”, Phonetica 50, (1993), pp. 180-188.

81. Lacheret-Dujour, A., Beaugendre, F., “La prosodie du français”, édition du CNRS, Paris, (1999).

82. Vannier, G., “Etude des contributions des structures textuelles et syntaxiques pour la prosodie: application à un système de synthèse vocale à partir du texte”, Thèse de Doctorat, Université de Caen, (1999).

83. Klatt, D.H., “Linguistic uses of segmental duration in English: acoustic and perception evidence”, Journal of the Acoustical Society of America, 59, (1976), pp. 1208-122.

84. Geoffrois, E., “Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole”, Thèse de Doctorat en sciences, Université de Paris XI Orsay, (1995).

85. Rossi, M., Di Cristo, A., Hirst, D., Martin, P., Nishinuma, Y., “L'intonation, de l'acoustique à la sémantique”, Klincksieck Press, Paris, (1981).

86. Calliope, La parole et son traitement automatique, Collection technique et scientifique des télécommunications, CNET 2 ENST, Edition Masson, 1989.

87. F. Beaugendre, Une étude perceptive de l'intonation du français, développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du texte, Thèse de doctorat en sciences de l'Université de Paris XI, Notes et Documents LIMSI n° 94-25, 1994.

88. A. Zaki, A. Rajouani, Z. Luxey, M. Najim, Rules Based Model for Automatic Synthesis of F0 Variation for Declarative Arabic Sentences, ISCA Archive, Speech Prosody 2002 Aix-en-Provence, France April 11-13, 2002

89. D. Kouloughli, Contribution à l’étude de l’accent en arabe littéraire, Annales de l’université d’Abidjan, série H, Vol. IX, pp.124-125, 1976.

90. E. Benveniste, Problèmes de linguistique générale, Paris, Gallimard, 1966.

91. P.A. Barbosa, Caractérisation et génération automatique de la structuration rythmique du français, Thèse de doctorat de l'Institut National Polytechnique de Grenoble, 1994.

92. J. Allen, M. S. Hunnicut, and D. H. Klatt, From Text to Speech: The MITalk system, Cambridge University Press, Cambridge, 1987.

93. D. O’Shaughnessy, A study of French vowel and consonant durations, Journal of phonetics, pp.385-406, 1981.

94. K. Bartkova, C. Sorin, A model of segmental duration for speech synthesis in French, Speech Communication, 6, M. Wajskop (éd.), Amsterdam, Elsevier, pp. 245-260, 1987.

95. Z. Zemirli, N. Vigouroux, Vers une modélisation de la durée des sons pour la génération automatique du rythme dans la synthèse de la langue arabe, actes des 23èmes Journées d’études sur la parole, Aussois, pp.261-264, 28-30 Juin 2000.

96. A. Amrouche, B. Boudraa et J.M Rouaven, Organisation temporelle des voyelles dans les structures CVCVCV, CVCCVCV, et CVCCV de l’arabe standard , Actes des 22èmes Journées d’études sur la parole, pp.91-94, 15-19 Juin 1998.

97. G. Droua Hamdani, Etude du paramètres durée de la prosodie de la langue arabe, Mémoire de Magister, CRSTDLA, Alger, Algérie, 2004.

98. J. V. Santen, Assignment of segmental duration in text-to-speech synthesis, Computer speech and language, 8:95-128, April 1994.

99. W.N. Campbell, syllable-based segmental duration, Edition G. Bailly and C. Benoît, Talking Machines: theories, Models and Designs, Elsevier Science Publishers, Amestrdam, pp.211-224, 1992.

100. H. Fujisaki, Physics of speech sounds, Tokyo University Press, 1967.

101. S. Ohman, Word and sentence intonation: a quantitative model, Quaterly Progress and Status Report, 2, K.T.H, Stockholm, 20-54, 1967.

102. J. Pierrehumbert, Synthesizing intonation, Journal of the Acoustical Society of America, 70, R.B. Lindsay (ed), New York, pp. 985-995, 1981.

103. D. Hirst, D. Espesser, Automatic modelling of fundamental frequency curves using a quadratic spline function, Travaux de l'Institut de Phonétique d'Aix, 15, pp. 71-85, 1993.

104. K. Silverman, M.E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labellinig English prosody, In ICLSP-92, volume 2, pages 867-870, 1992.

105. P. Mertens, L'intonation du français. De la description linguistique à la reconnaissance automatique, Doctorale dissertatie, Katholieke Universiteit Leuven, Faculteit van de Letteren en Wijsbegeerte, Departement Linguïstiek, 1987.

106. J. Pierrehumbert, The phonology and phonetics of English intonation, Doctoral dissertation, Massachussets Institute of Technology, Cambridge, MA, 1980.

107. A. Cruttenden, Intonation, Cambridge University Press, Cambridge, 1986.

108. P. Delattre, Les dix intonations de base du français, French Review, 40, American Association of Teachers of French, Illinois, 1-14, 1966.

109. A. Di Cristo, De la microprosodie à l'intonosyntaxe, thèse de Doctorat d'état, Université de Provence, Aix-Marseille I, 1978.

110. P. Martin, Prosodic and rythmic structures in French, Linguistics 25, p.925-949, Mouton de Gruyter, Amsterdam, 1987.

111. V. Aubergé, La synthèse de la parole : des règles aux lexiques, Thèse de Doctorat d'informatique, Université Pierre Mendès France, Grenoble, 1991.

112. C. Trabe, F_0 generation with a database of natural F_0 patterns and with a neural network, Editions G. Bailly et C. Benoît, Talking Machines : Theories, Models and Designs, Elsevier Science Publishers, B.V., pp.287-304, 1992.

113. D. Hirst, J. Véronis, N. Ide, Analysis of fundamental frequency patterns for multi-lingual synthesis using INTSINT, Proceeding of 2nd ESCA/IEEE Workshop on Speech Synthesis, New York, September 1994.

114. O. Boëffard, Segmentation automatique d'unités acoustiques pour la synthèse de la parole, Thèse de Doctorat, Université de Rennes I, 1993.

115. A.W. Black & A.J. Hunt, Generating F_0 contours from ToBI labels using linear regression, Proceedings of ICSLP'96, pp. 1385-1388, 1996.

116. B. Möbius, M. Pätzold & W. Hess, Analysis and synthesis of German F_0 contours by means of Fujisaki's Model, Speech Communication, pp. 53-61, 1993.

117. L. Es-Skali, A. Rajouani, M. Najim, D. Chidami, Elements d'un modèle intonatif pour la phrase affirmative en arabe. In Actes des XVI^{ème} JEP, Hammamet, pp.282-285, 1987.

118. A. Rajouani, Contribution à la synthèse de la parole arabe par règles, Thèse de Doctorat d'état, Université Mohamed V, Faculté des sciences Rabat, 1989.

119. D. Hirst, A. Di Cristo , Intonation Systems. A Survey of Twenty Languages, Cambridge University Press, 1998.

120. Vladimir Goncharoff and Patrick Gries, An Algorithm for accurately marking pitch pulses in speech signals, Proceedings of the IASTED International Conference, Signal and image processing (SIP'98), pages 281-234, Las Vegas, Nevada, USA, October 28-31,1998.

121. Y. M. Cheng and D. O'shaughnessy, Automatic and reliable estimation of glottal closure instant and period, IEEE Trans. Acoust., Speech, Signal Processing, ASSP-37(12):1805-1815, December 1989.

122. Y. Laprie and V. Colotte. Automatic pitch marking for speech transformation via TD-PSOLA. In IX European Signal Processing Conference, Rhodes, Greece, 1998.

123. F.M. Giménez de los Galanes, M.H. Savoji and J.M. Pardo, New algorithm for spectral smoothing and enveloppe modification for LP-PSOLA synthesis, IEEE, 1994.

124. M. Edgington and A. Lowry, Residual-Based Speech modification algorithm for text-to-speech synthesis, ICLSP'96, Philadelphia, PA, USA, pp. 1425-1428, October 3-6, 1996.

125. Y. Medan, E. Yair et D. Chazan, Super resolution pitch determination of speech Signals, IEEE Trans. Acoust., Speech, Signal processing 39(1), p. 40-48, 1991.

126. M. Oudot, Etude du modèle Sinusoïdes et Bruit pour le traitement des signaux de parole, Estimation robuste de l'enveloppe spectrale, Thèse de doctorat de l'ENST Paris, France, Novembre 1998.

127. B. Doval, Estimation de la fréquence fondamentale des signaux sonores, Université Paris, VI. Thèse de Doctorat, 1994.

128. D. W. Griffin and J.S. Lim. Multiband excitation vocoder. In IEEE Trans. Acoust., Speech, signal Processing, volume 36, pages 1223-1235, august 1988.

129. Y. Stylianou, Decomposition of speech signal into a deterministic and a stochastic part, Proc. of ICSLP 1996.

130. S. Kay, Modern spectral estimation, Prentice Hall, Englewood cliffs, New Jersey, 1988.

131. T. Galas and X. Rodet, An improved cepstral method for deconvolution of source-filter systems with discret spectra: application to musical sound signals, In ICMC, 1990.

132. O. Cappé and E. Moulines, Regularization techniques for discrete cepstrum estimation. IEEE Signal Processing Letters, 3(4): 100-102, apr 1996.

133. O. Cappé, J. Laroche, and E. Moulines, Regularized estimation of cepstrum envelope from discrete frequency points, In IEEE ASSP Workshop on App. Of Sig. Proc. To Audio and Acoust., Mohonk, October 1995.

134. M. Boudraa and B. Boudraa, Twenty Lists of Ten Arabic Sentences for Assessment, Acustica. Acta Acustica, 86, pp. 870–882, 2000.

135. F. Chouireb and M. Guerti, Extraction des paramètres prosodiques de l'Arabe Standard, International Conference on Control, Modelling and Diagnosis, ICCMD'06, Annaba Algeria, pp. 67–71, May 22-24, 2006.

136. F. Chouireb, M. Guerti, M. Naïl, and Y. Dimeh, Development of a Prosodic Database for Standard Arabic, The Arabian Journal for Science and Engineering, Volume 32, Number 2B, pp. 251-262, ISSN: 1319-8025, October 2007.

137. F. Malfrère, O. Deroo, and T. Dutoit, Phonetic Alignment: Speech Synthesis Based vs. Hybrid HMM/ANN, Proceedings of ICSLP 98, Sydney, 1998.

138. F. Malfrère and T. Dutoit, Speech Synthesis for Text- to-Speech Alignment and Prosodic Feature Extraction, Proceedings of ISCAS' 97, 1997.

139. T. Dutoit, V. Pagel, N. Pierret, F. Bataille and O. vander Vrecken, The MBROLA Project : Towards a Set of High Quality Speech Synthesizers Free of Use for Noncommercial Purposes, International Conference on Speech and Language Processing, Philadelphia, 1996.

140. J. Nouza , Spectral Variation Functions Applied to Acoustic-Phonetic Segmentation of Speech Signal, In: HW Wodarz Ed, Speech Processing (Forum Phoneticum, 63), Frankfurt amndt Hand, pp. 43–58, 1997.

141. Laurent Buniet, Traitement automatique de la parole en milieu bruité : Etude de modèles connexionistes statiques et dynamiques, Thèse de Doctorat, Université Henri poincaré-Nancy 1, 1997.

142. T. Kohonen, K. Mäkisara & T. Saramäki, Phonotopic maps - Insightful representation of phonological features for speech recognition, Proceedings of the IEEE International Conference on Pattern Recognition, pp 182-185, 1984.

143. T. Kohonen, Self-organisation and associative memory. 312 pp, Springer series in information sciences, vol. 8, Springer-Verlag, Berlin (Allemagne), 2ème édition, 1987.

144. Y. Le Cun, Une procédure d'apprentissage pour réseau à seuil asymétrique, Actes de la conférence Cognitiva, pp 599-604, 1985.

145. C. TOUZET, Les réseaux de neurones artificiels : introduction au connexionnisme, 150 pages, Préface de Jeanny Hérault, EC2 éd., Paris, 1992.

146. F. Itakura, Line spectrum representation of linear prediction coefficients of speech signals, Journal Acoustical Society of America, vol. 57, pp. 535. (abstract), 1975.

147. W. B. Kleijn and K. K. Paliwal, (Eds.), Speech Coding and Synthesis, Amsterdam: Elsevier, 1995.

148. J. Vepa, S. King, Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis, IEEE Trans. On Speech and Audio Processing, Sept. 2006, Volume 14, Issue 5, pp. 1763 - 1771.

149. Hirst, Daniel, Albert Di Christo, and Robert Espesser, Levels of representation and levels of analysis for intonation, in M. Horne (ed) Prosody: Theory and Experiement, Kluwer academic Publishers, Dordrecht, pp.51-87, 2000.

150. G. Fant, and A. Kruckenberg, Intonation analysis and synthesis with reference to Swedish, International Somposium on Tonal Aspects of languages: with Emphasis on Tone languages, Beijing, China, March 28-31, 2004.

151. J. Hart, R. Collier, and A. Cohen, A perceptual study of intonation, Cambridge University Press, Cambridge, 1990.

152. A. Conkie and S. Isard, Optimal coupling of diphones, In J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg,editors, Progress in speech synthesis, pages 293–304. Springer-Verlag, New York, pages 293–304, 1997.

153. A. Black and N. Campbell, Optimising selection of units from speech databases for concatenative synthesis, In EUROSPEECH '95, Madrid, Spain, pages 581-584, 1995.

154. John H. L. Hansen and David T. Chappell, An Auditory-Based Distortion Measure with Application to Concatenative Speech Synthesis, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 5, pp. 489-495, September 1998.

155. T. Dutoit and M. Cernak, TTSBOX: A MATLAB toolbox for teaching text-to-speech synthesis, ICASSP'05, Philadelphia, pp: v/537 - v/540 Vol. 5, 18-23 March 2005.