Synthèse paramétrique de la parole Arabe

Amal Houidhek

Ecole Doctorale IAEM (Informatique, Automatique, Electronique-Electrotechnique, Mathématiques et Sciences de l'Architecture)

Thèse

PRESENTEE ET SOUTENUE PUBLIQUEMENT POUR L'OBTENTION DU TITRE DE

DOCTEUR DE L'UNIVERSITE DE LORRAINE

MENTION : « INFORMATIQUE »

Par Amal HOUIDHEK

Synthèse paramétrique de la parole Arabe

12 Février 2020

Membres du jury :

M. Zied Laachiri	**Professeur, Ecole Nationale d'ingénieurs de Tunis, Tunis**
M. Adnen Cherif	**Professeur, Faculté des sciences de Tunis, Tunis**
M. Yannick Estève	**Professeur à l'Université d'Avignon, Avignon**
M. Hamid Amiri	**Professeur Émérite, Ecole Nationale d'ingénieurs de Tunis, Tunis (Directeur de thèse)**
M. Denis Jouvet	**Directeur de recherche, Loria-Inria, Nancy (Directeur de thèse)**
M. Vincent Colotte	**Maître de conférences à l'Université de Lorraine, Nancy (Co-Directeur de thèse)**

Ecole Doctorale IAEM (Informatique, Automatique, Electronique-Electrotechnique, Mathématiques et Sciences de l'Architecture)

Thèse

PRESENTEE ET SOUTENUE PUBLIQUEMENT POUR L'OBTENTION DU TITRE DE

DOCTEUR DE L'UNIVERSITE DE LORRAINE

MENTION : « INFORMATIQUE »

Par Amal HOUIDHEK

Synthèse paramétrique de la parole Arabe

12 Février 2020

Membres du jury :

M. Zied Laachiri	**Professeur, Ecole Nationale d'ingénieurs de Tunis, Tunis**
M. Adnen Cherif	**Professeur, Faculté des sciences de Tunis, Tunis**
M. Yannick Estève	**Professeur à l'Université d'Avignon, Avignon**
M. Hamid Amiri	**Professeur Émérite, Ecole Nationale d'ingénieurs de Tunis, Tunis (Directeur de thèse)**
M. Denis Jouvet	**Directeur de recherche, Loria-Inria, Nancy (Directeur de thèse)**
M. Vincent Colotte	**Maître de conférences à l'Université de Lorraine, Nancy (Co-Directeur de thèse)**

Résumé

Cette thèse porte sur l'adaptation de la synthèse paramétrique de la parole à partir d'un texte écrit à la langue arabe. Pour ce faire, différentes méthodes ont été développées afin de mettre en place des systèmes de synthèse. Ces méthodes sont basées sur une description du signal de parole par un ensemble de paramètres acoustiques et prosodiques. De même, chaque son est représenté par un ensemble de descripteurs contextuels contenant toutes les informations affectant la prononciation de celui-ci.

Une partie de ces descripteurs dépend de la langue et de ses particularités, ainsi, afin d'adapter l'approche de synthèse paramétrique à l'arabe, une étude des particularités phonologiques de l'arabe était nécessaire. L'accent a été mis sur deux phénomènes : la gémination et la longueur des voyelles (courte/longue). Deux descripteurs associés à ces deux phénomènes ont été ajoutés à l'ensemble des descripteurs contextuels. De même, différentes approches de choix des unités ont été proposées pour modéliser les consonnes géminées et les voyelles longues. Quatre combinaisons de modélisation sont possibles en alternant la différentiation ou la fusion des consonnes simples et géminées d'une part et des voyelles courtes et longues d'autres part.

Un ensemble des tests perceptifs et objectifs a été conduit afin d'évaluer l'effet des quatre approches de modélisation des unités sur la qualité de la parole synthétisée. Les évaluations ont été faites dans le cas de synthèse paramétrique par HMM (Hidden Markov Model) puis dans le cas de la synthèse paramétrique par DNN. Les résultats subjectifs ont montré que dans le cas de l'approche par HMM, les quatre approches produisent des signaux de qualité similaire, une conclusion qui a été confirmée par les mesures objectives calculées pour évaluer la prédiction des durées des unités de parole. Cependant, les résultats des évaluations objectives dans le cas de l'approche par DNN ont montré que la différentiation des consonnes simples (respectivement des voyelles courtes) des consonnes géminées (respectivement des voyelles longues) permet d'avoir une prédiction des durées légèrement meilleure qu'avec les autres des approches de modélisation. En revanche, cette

amélioration n'a pas été perçue lors des tests perceptifs ; les participants ont trouvé que les signaux générés par les quatre approches sont similaires en termes de qualité globale. Une dernière partie de la thèse a été consacrée à la comparaison de l'approche de synthèse par HMM à celle par DNN. L'ensemble des tests conduits ont montré que l'utilisation des DNN a amélioré la qualité perçue des signaux générés.

Abstract

The presented thesis deals with the adaptation of the conversion of a written text into speech using a parametric approach to the Arabic language. Different methods have been developed in order to set up synthesis systems. These methods are based on a description of the speech signal by a set of parameters. Besides, each sound is represented by a set of contextual features containing all the information affecting the pronunciation of this sound.

Part of these features depend on the language and its peculiarities, so in order to adapt the parametric synthesis approach to Arabic, a study of its phonological peculiarities was needed. Two phenomena were identified : the gemination and the vowels quantity (short / long). Two features associated to these phenomena have been added to the contextual features set. In the same way, different approaches have been proposed to model the geminated consonants and the long vowels of the speech units. Four combinations of modeling are possible : alternating the differentiation or fusion of simple and geminated consonants on the one hand and short and long vowels on the other hand.

A set of perceptual and objective tests was conducted to evaluate the effect of the four unit modelling approaches on the quality of the generated speech. The evaluations were made in the case of parametric synthesis by HMM then in the case of parametric synthesis by DNN. The subjective results showed that when the HMM approach is used, the four approaches produce signals with a similar quality, this result that was confirmed by the objective measures calculated to evaluate the prediction of the durations of the speech units. However, the results of objective evaluations in the case of the DNN approach have shown that the differentiation of simple consonants (respectively short vowels) geminated consonants (respectively long vowels) leads to a slightly better prediction of the durations than the other modelling approaches. On the other hand, this improvement was not perceived during the perceptive tests ; listeners found that the signals generated by the four approaches are similar in terms of overall quality. The last part of this thesis was

devoted to the comparison of the synthesis approach by the HMMs to that by the DNNs. All the tests conducted have shown that the use of DNNs has improved the perceived quality of the generated signals.

Remerciements

Ce travail de recherche a été réalisé dans le cadre du programme PHC-Utique dans le cadre de la subvention CMCU (Comité Mixte de Coopération Universitaire) sous le numéro 15G1405.

Il me sera très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement mon directeur de thèse, Monsieur Denis Jouvet, Directeur de Recherche à l'INRIA pour toute son aide. Je suis ravi d'avoir travaillé en sa compagnie car outre son appui scientifique, il a toujours été là pour me soutenir et me conseiller au cours de l'élaboration de cette thèse.

Je tiens à remercier Monsieur Vincent Colotte, Maître de conférences à l'Université de Lorraine, qui m'a encadré tout au long de cette thèse et qui m'a fait partager ses brillantes intuitions. Qu'il soit aussi remercié pour sa gentillesse, sa disponibilité permanente et pour les nombreux encouragements qu'il m'a prodiguée. Je remercie mon directeur de thèse, Monsieur Hamid Amiri, Professeur émérite à l'École Nationale d'Ingénieurs de Tunis. Cette thèse est le fruit d'une collaboration de plus de cinq années avec lui.

Monsieur Zied Mnasri, Maître assistant à l'École Nationale d'Ingénieurs de Tunis, m'a initié à la synthèse de parole et m'a aussi prodigué de nombreux conseils pour bien débuter le troisième cycle universitaire dont cette thèse est l'accomplissement.

Je tiens à remercier le jury pour avoir jugé et analysé mes travaux. Un grand merci à Zied Laachiri pour avoir présidé mon jury.

Merci à Monsieur Adnen Cherif Professeur à la Faculté des Sciences de Tunis et Monsieur Yannick Estève, Professeur à l'Université d'Avignon, pour m'avoir fait l'honneur d'être rapporteurs de ma thèse et pour avoir également rapporté mes travaux.

Table des matières

Table des figures

Liste des tableaux

Glossaires

ACR Absolute Category Rating 23, 24

BLSTM Bidirectional Long Short Term Memory 55, 57, 102, 103

BPTT Back Propagation Through Time 55

CCR Comparaison Category Rating 24

DCR Degradation Category Rating 24

DMOS Differential Mean Opinion Score IX, X, XII, 24, 68, 81, 91, 92, 95, 97, 111, 112, 118, 133

DNN Deep Neurel Network II–IV, VIII, XII, 2–4, 19, 27, 28, 50, 53–57, 79, 95, 96, 98–101, 103, 106, 108–115, 117–119

GV Global Variance 50

HMM Hidden Markov Model II–IV, VIII, XI, 2–4, 19, 27, 28, 35–37, 39, 41, 42, 44–48, 50, 54–57, 69, 79, 81–83, 90, 94–96, 98, 99, 108–115, 117–119

HTS HMM-based Speech Synthesis System ou H Triple S XI, 2–4, 36, 37, 39, 48, 49, 57, 68, 69, 77, 78, 81–83, 85, 88, 89, 91, 93, 95, 97, 99, 100, 108, 110, 118

IPA International Phonetic Alphabet 78

LSTM Long Short Term Memory 55–57, 102

MDL Minimum Description Length 39

MLPG Maximum Likelihood Parameter Generation XI, 42, 43, 83

MLSA Mel-Log Spectrum Approximation 31, 37, 83

MOS Mean Opinion Score X, 23–25, 56, 68–70, 89, 90, 95, 110, 111, 118, 131

MSA Modern Standard Arabic 2, 60, 61, 63–65

MSD Multi-Space Probability Distribution XI, 44, 45

MUSHRA MUltiple phrase with Hidden Reference and Anchor 25

NRMSE Normalized Root Mean Square Error XII, XIII, 87, 88, 104, 105, 109, 110

PESQ Perceptual Evaluation of Speech Quality 22, 23

POLQA Perceptual Objective Listening Quality Analysis 22

PSOLA Pitch Synchronous Overlap Add 17

RMSE Root Mean Sqaure Error XII, XIII, 21, 56, 87, 101, 103–105, 109, 110

RNN Recurrent Neural Network 55–57

STRAIGHT Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum 30, 31, 36, 57, 69, 100

TD-PSOLA Time-Domain Pitch-Synchronous Overlap-and-Add 17

TTS Text To Speech XI, 13, 19

Introduction générale

L'intérêt a été porté sur la production artificielle de la parole depuis longtemps. Le but des différents systèmes implémentés a été la production d'une parole naturelle et intelligible. Une première machine de production de parole artificielle a été conçue par Von Kempelen en 1791. Elle reposait sur une imitation de la physiologie de l'appareil phonatoire. Il s'agissait d'une sorte d'instrument à vent constitué de divers étages destinés à reproduire les organes humains entrant en jeu lors de la production de parole. Les travaux se sont poursuivis jusqu'à obtenir des systèmes de synthèse reposant sur des approches de conversion des textes en des signaux de parole sont apparues (Taylor, 2009). Ces méthodes sont basées sur l'extraction de certaines informations à partir du texte afin de le convertir en un signal de parole.

Le tout premier système de cette génération est basé sur l'utilisation des règles qui sont déduites après l'observation des spectrogrammes de la parole naturelle (prononcée par un être humain) (Klatt, 1980). Ces règles décrivent la prononciation des unités de parole et l'évolution temporelle des formants ce qui permet de générer un spectre de signal de parole. Ensuite, des corpus contenant des unités pré-enregistrées de parole de différentes tailles ont été utilisées (1 exemplaire de chaque unité) dans la synthèse de parole par concaténation c'est-à-dire par une simple mise bout à bout des segments de parole (Moulines *et al.*, 1990). Ces méthodes de concaténation ont évolué avec le développement des calculateurs et les mémoires de stockage ; chaque unité de parole du corpus est enregistrée dans différents contextes phonétiques et prosodiques. La synthèse de parole est alors réalisée par une sélection de la meilleure séquence d'unités de parole correspondantes au texte (Hunt et Black, 1996). Ceci en se basant sur deux critères qui sont un coût de cible pour mesurer la similarité entre les caractéristiques des unités sélectionnées et les cibles désirées, et un coût de concaténation pour mesurer la qualité de la concaténation des unités sélectionnées.

L'utilisation des modèles statistiques a permis une évolution majeure de la synthèse de parole pendant les années 90s du siècle précédent et a permis la mise en œuvre de la

synthèse de parole par approche paramétrique statistique (Black *et al.*, 2007). Les modèles de Markov cachées HMM ont été utilisés dans cette approche ; un emploi justifié par la robustesse que ces modèles ont montré lors de leur utilisation dans la reconnaissance de la parole. Le système de synthèse résultant est dénommé HTS (HMM-based Speech Synthesis System ou H Triple S) a été largement utilisé et adapté à plusieurs langues. Récemment, les HMM ont été remplacé par les DNN (Deep Neurel Network) dans l'approche de synthèse paramétrique et ceci a permis d'améliorer la qualité de la parole synthétisée (Zen, 2013).

Dans ce présent travail, l'attention est portée à la langue arabe plus précisément le MSA (Modern Standard Arabic) (Al-Ani, 1970). Il s'agit d'une langue sémitique dont le nombre de locuteurs est estimé à 375 millions de personnes. La langue arabe présente certaines particularités phonologiques et phonétiques, à savoir la gémination, les voyelles longues et l'aspect emphatique de certaines consonnes (Newman, 2002). Les travaux décrits dans cette thèse se sont particulièrement intéressés à la synthèse par approche paramétrique. Cette méthode exige une description de chaque son du texte à synthétiser par un ensemble d'informations de différents types (linguistiques, phonologiques, prosodiques...) appelés descripteurs contextuels (Tokuda *et al.*, 2002). Ces informations permettent de différentier les différents contextes dans lesquels une unité de parole (phonème) peut exister. Une partie des descripteurs dépend de la langue, ainsi, afin d'appliquer l'approche paramétrique de synthèse à une langue, il faut commencer par adapter l'ensemble des descripteurs aux particularités de cette langue.

Cette thèse n'est pas le premier travail qui s'intéresse à la synthèse de la parole arabe ; différentes approches ont été adaptées à la langue arabe précédemment. Entre autres, l'approche paramétrique statistique a été utilisée pour la génération de signaux en arabe sans porter un intérêt explicite aux particularités de la langue. En revanche, le but des travaux de cette thèse est d'appliquer l'approche de synthèse paramétrique statistiques (HMM) à la langue arabe tout en considérant les phénomènes de gémination et les voyelles longues. L'ensemble de descripteurs a été adapté aux particularités de la langue arabe en ajoutant des descripteurs faisant référence à la gémination et la longueur des voyelles. A ce niveau, on s'est posé différentes questions : Comment modéliser les consonnes géminées et les voyelles longues ? Doit-on les modéliser avec les mêmes modèles que les consonnes simples et les voyelles courtes ou avec des modèles différents ? Ainsi, nous avons proposées quatre approches différentes de modélisation pour faire la synthèse de la parole arabe. Les signaux synthétisés ont ensuite été évalués par un ensemble des tests objectifs et subjectifs.

Dans un second temps, les HMM ont été remplacées par des DNN. Le même processus que pour la synthèse par HMM a été suivi et les quatre approches de modélisation ont été utilisées pour générer des signaux en arabe qui ont été évaluées. Finalement, une comparaison entre l'utilisation des HMM et des DNN a eu lieu. L'intérêt a été porté sur l'impact de chaque approche sur la qualité des signaux produits.

Afin de présenter les travaux réalisés, ce document est divisé en cinq chapitres. Un état de l'art concernant la parole et les différentes approches de synthèses citées dans la littérature est présenté dans le chapitre 1. Une première section décrit l'aspect physiologique de la parole, les caractéristiques du signal de parole humain ainsi que la manière de modéliser mathématiquement un signal de parole. Une deuxième section est consacrée à la description du processus de synthèse de la parole et des différents systèmes implémentés qui sont basés sur différentes approches. Une dernière section est dédiée aux différentes méthodes d'évaluation des systèmes de synthèse de parole : les mesures objectives et les tests perceptifs.

Le deuxième chapitre est consacré à une description détaillée de la synthèse de parole par approche paramétrique. La première section décrit l'aspect général de l'approche et sa particularité par rapport aux autres méthodes décrites dans le chapitre précédent. La section suivante s'intéresse à l'approche de synthèse paramétrique statistique par HMM : une description de l'aspect général, la motivation d'utilisation des HMM et la modélisation des paramètres acoustiques. Cette section décrit aussi le toolkit HTS utilisé pour les expériences. La dernière section décrit l'approche de synthèse paramétrique par DNN. Cette partie commence tout d'abord par décrire les avantages qu'apporte l'introduction des DNN dans la synthèse paramétrique de parole et se poursuit par une présentation de l'aspect général du système de synthèse par les DNN. Ce chapitre est clôturé par une description du toolkit MERLIN qui a été utilisé pour générer des signaux par DNN.

Une première partie du chapitre 3 "Adaptation de la Synthèse paramétrique à la parole arabe", est dédiée à la description des travaux qui ont été déjà faits sur la synthèse de la parole arabe en utilisant les différentes approches de synthèse décrites dans le chapitre 1. La suite du chapitre correspond au travail de thèse et traite les approches de modélisation des unités de parole ainsi que l'ensemble des descripteurs contextuels proposés afin d'appliquer l'approche de synthèse paramétrique à la langue arabe. Le chapitre se termine par une description des données expérimentales utilisées pour les expériences.

Le chapitre 4 "Synthèse de la parole par HMM" décrit les expériences de synthèse de

la parole arabe par HMM en utilisant le toolkit HTS. Une première section rappelle le principe de synthèse par approche paramétrique statistique (par HMM) suivie par la mise en oeuvre de HTS pour la langue arabe. Les deux sections suivantes présentent les résultats de l'évaluation objective des durées prédites par les quatre approches de modélisation ainsi qu'une évaluation subjective des signaux générés avec les quatre approches.

Le dernier chapitre (chapitre 5) "Synthèse de la parole arabe par DNN" décrit les expériences réalisées pour faire la synthèse de la parole arabe en utilisant les DNN. Pareillement au chapitre précédent, le chapitre commence par un rappel de l'aspect général de la synthèse de la parole par DNN. Puis la mise en oeuvre du toolkit MERLIN pour la synthèse de la parole arabe. Les deux sections suivantes décrivent les résultats des évaluations faites : tout d'abord une évaluation objective des durées prédites et des paramètres acoustiques générés des signaux synthétisés, ceci en utilisant les quatre approches de modélisation. Par la suite, une évaluation subjective des quatre approches de modélisation a été conduite. Cette partie finit par une comparaison des performances des HMM et des DNN.

Chapitre 1

Synthèse de la parole

Sommaire

1.1 Introduction

La parole est le premier moyen de communication entre les hommes. De nos jours, la communication homme-machine devient de plus en plus répandue, et le principe a

été inspiré de la communication entre les hommes, plus particulièrement du phénomène naturel de la production et la perception de la parole afin de rendre la communication la plus proche possible de la parole naturelle. Ce chapitre présente l'état de l'art de la synthèse de la parole. Une première partie s'intéresse à la parole : son aspect physiologique, les caractéristiques du signal de parole ainsi que sa modélisation acoustique. La section suivante introduit quelques notions de phonétiques dont la compréhension est nécessaire à la synthèse de la parole. Ensuite, une présentation de la synthèse de parole à travers la description du principe de fonctionnement et les différentes approches utilisées. La dernière section de ce chapitre décrit les méthodes d'évaluation des systèmes de synthèse de la parole.

1.2 La parole

Cette partie décrit en premier lieu l'interaction entre les systèmes physiologiques qui est à l'origine de la production de la parole. Cette partie sera suivie par une présentation des caractéristiques du signal de parole produit et sa modélisation source-filtre.

1.2.1 Aspect physiologique

Le cerveau envoie les commandes de contrôle des mouvements nécessaires afin de produire la parole. Le processus de la génération de la parole est un mécanisme complexe assuré par l'appareil phonatoire (représenté dans la figure 1.1) qui fait intervenir plusieurs organes tels que les poumons, le larynx, la langue et les lèvres (Divenyi *et al.*, 2006).

Le fonctionnement de l'appareil phonatoire repose sur une interaction à trois niveaux (Vilain, 2002), (Kamina, 2014) :

Niveau respiratoire : appelé aussi système subglottique. C'est la source de souffle qui est assurée par les poumons, le diaphragme et la trachée.

Niveau glottique : appelé aussi système phonatoire ou larynx. C'est la source vocale ou sonore. Elle comprend les cordes vocales, des cartilages et des muscles.

Niveau articulatoire : appelé aussi système supra-glottique. Il contient les résonateurs qui sont principalement les cavités pharyngale, buccale, nasale et labiale.

Le phénomène de production de la parole résulte d'une variation de la pression de l'air générée par l'appareil phonatoire.

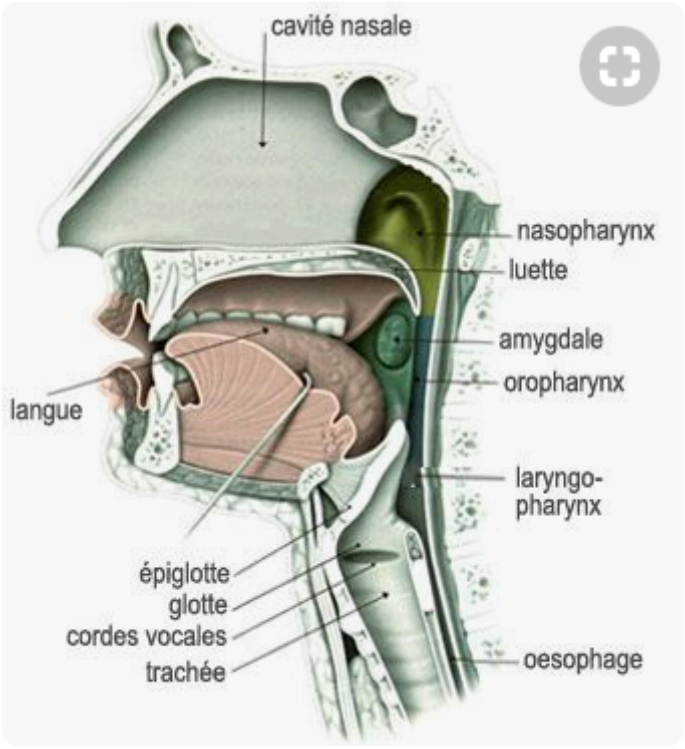


FIGURE 1.1 – L'appareil phonatoire.[1]

Le mécanisme commence par un souffle : les poumons se gonflent et se dégonflent afin d'entretenir un courant d'air qui atteint les cordes vocales. Les cordes vocales sont définies par la superposition des muscles et de ligaments (figure 1.2). Un bruit est produit si les cordes vocales sont écartées pour laisser passer l'air librement. Dans le cas contraire, quand les cordes vocales sont rapprochées, l'air sous pression peut engendrer leur vibration ce qui mène à avoir un son quasi-périodique dont la fréquence fondamentale correspond à la hauteur de la voix perçue (En-Najjary, 2005).

Les cordes vocales possèdent trois positions possibles :

— Cordes écartées : l'air circule librement

— Cordes accolées : l'air ne passe pas

— Cordes rapprochées et vibrant : l'air circule en faisant vibrer les cordes vocales, c'est le phénomène de voisement.

1. Source : (Kamina, 2014)

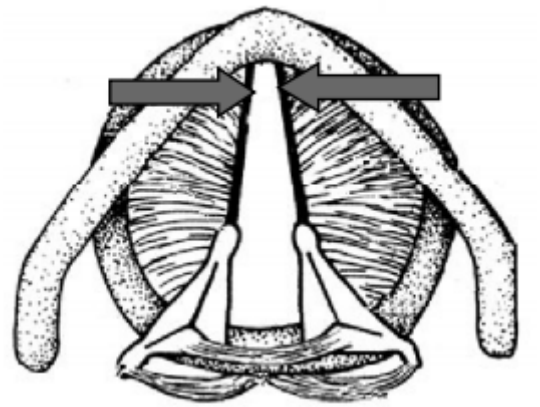


FIGURE 1.2 – Coupe d'une corde vocale

Dans les cas où les cordes vocales sont écartées ou rapprochées, le flux d'air continue son chemin à travers le conduit vocal. Ce dernier est considéré comme résonnateur de parole. Sa forme est déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du palais ce qui définit le timbre des différents sons de la parole.

1.2.2 Caractéristiques du signal de la parole

Le signal de parole obtenu est un signal continu périodique ou aléatoire. Ainsi, le signal de parole est caractérisé par :

Le voisement Un signal de parole est formé par une succession des sons voisés et des sons non voisés dont l'amplitude diffère :

Les sons voisés : Ils sont considérés comme des signaux quasi-périodiques ayant une fréquence fondamentale et des harmoniques. Ils correspondent aux voyelles (/a/, /i/, /o/, /u/), des consonnes (/m/, /n/, /b/, /d/), et ce sont des sons prononcés avec une vibration des cordes vocales.

Des sons non voisés : Ils correspondent à un écoulement d'air turbulent. Ainsi, ces sons sont généralement considérés comme des bruits fricatifs.

La fréquence fondamentale : La fréquence fondamentale d'un signal de parole est la fréquence d'oscillation (quasi-périodiques) des cordes vocales résultant de l'écoulement d'air des poumons. Il s'agit donc d'une mesure physique. Généralement le terme

"pitch" est utilisée comme synonyme de la fréquence fondamentale, même si le pitch fait référence à la fréquence de la tonalité perçue, c'est ce que l'être humain peut entendre. Ainsi, la fréquence fondamentale est un paramètre acoustique mais le pitch reste un paramètre de perception. Généralement, la fréquence fondamentale varie selon le genre et l'âge :

- homme de 100 à 150 Hz
- femme de 200 à 300 Hz
- enfant de 300 à 450 Hz

L'énergie : Ce paramètre caractérise l'intensité sonore d'un segment de parole.

Le spectre fréquentiel : Le signal de parole peut être représenté dans le domaine temporel et dans le domaine fréquentiel par respectivement la forme d'onde et le spectre fréquentiel. Sur des intervalles courts, le principe de Fourier permet de décomposer le signal de parole en une somme d'ondes sinusoïdales. L'analyse de Fourier permet de déterminer quelles sinusoïdes sont à considérer pour reconstruire le signal original. Les amplitudes de ces ondes sinusoïdales révèlent le contenu fréquentiel du signal d'origine.

1.2.3 Modélisation du signal de la parole

Pour faire une représentation numérique du signal de parole, une modélisation source-filtre peut être utilisée (Taylor, 2009). Cette théorie source-filtre permet de décrire le processus de production de la parole en termes de deux contributions indépendantes : la source sonore et le filtre du canal vocal. Deux hypothèses sont à l'origine de cette modélisation :

- La modulation du flux d'air passant par la glotte est indépendante des variations du conduit vocal. L'indépendance de la source et du filtre signifie que chacun apporte une contribution distincte aux caractéristiques des sons de la parole produite. La source est responsable de l'amplitude du voisement tandis que le filtre du système vocal est responsable de la localisation des formants et de la forme spectrale globale.
- Le signal de parole peut être décrit comme étant le résultat de la convolution de l'onde produite par la vibration des cordes vocales (la source) par les résonateurs du conduit vocal (filtre) (figure 1.3).

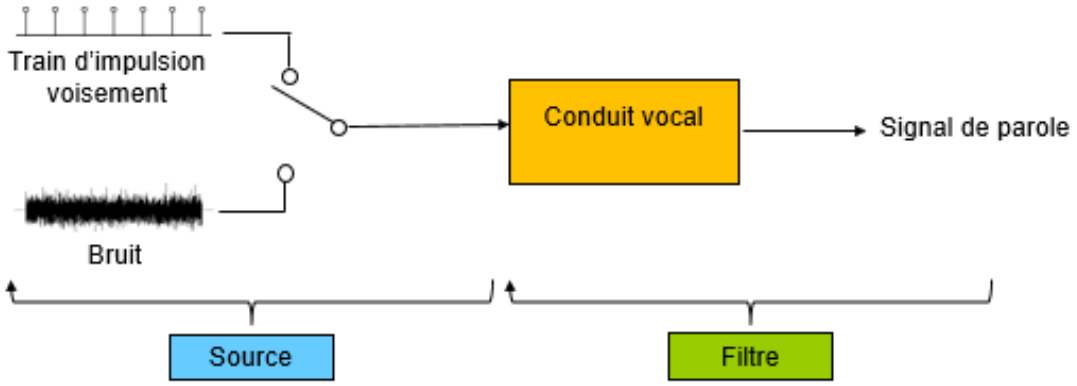


FIGURE 1.3 – Modélisation source/filtre

La phase de filtrage correspond à l'amplification et au filtrage fréquentiel de l'onde source produite par la glotte. Ces fonctions sont reprises de manière complexe par les organes supra glottiques qui renforcent ou atténuent certaines fréquences (figure 1.4).

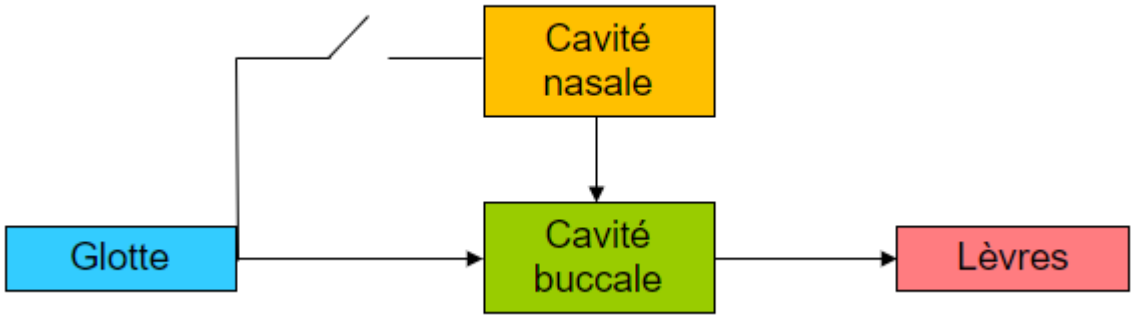


FIGURE 1.4 – Conduit vocal

Ces résonateurs sont des cavités de forme et de taille variables, ce qui permet d'ajuster le timbre du son, via les phénomènes de résonance. Les différentes cavités du conduit vocal vont servir de résonateur au signal de la source. Généralement, chaque cavité de résonance est caractérisée par une fréquence de résonance propre qui dépend de certains facteurs, telle que la longueur de la cavité. Le principe de cette modélisation est utilisé dans les

systèmes de synthèse paramétrique ainsi que dans les vocodeurs.

1.3 Quelques notions de phonétique

La plus petite unité de parole produite est le phonème. Il s'agit de la plus petite unité qui, si elle est modifiée, change le sens du mot. Les phonèmes sont classés en deux familles ; les consonnes et les voyelles (Lucci, 1983). La distinction en consonnes et voyelles se fait comme suit :

— Une voyelle est produite à la suite d'un passage d'air librement dans la glotte.

— Une consonne est produite à la suite d'un passage obstrué de l'air en un ou plusieurs endroits.

Généralement, le timbre des voyelles est étroitement lié aux résonateurs du conduit vocal ; leur nombre, la forme du résonateur buccal et son volume. Les voyelles peuvent être regroupées en différentes classes selon les critères suivants :

— La zone d'articulation.

— La nasalité.

— La forme des lèvres.

— La forme du conduit vocal.

En outre, les consonnes peuvent être classées selon le mode et le point d'articulation.

Le mode d'articulation :

Il s'agit de l'ensemble des propriétés de son articulation qui modifie la nature du courant d'air expiré. Deux modes d'articulation consonantique sont possibles :

— Un passage d'air fermé pendant un court instant permet de produire des consonnes occlusives (exemple : /b/, /d/).

— Un passage d'air rétrécit produit des consonnes continues ou fricatives (exemple : /f/, /s/).

Le point d'articulation :

Ce point fait référence à l'organe articulateur (lèvres, langues, ...). Les consonnes produites selon chaque point d'articulation sont regroupées dans la figure 1.5.

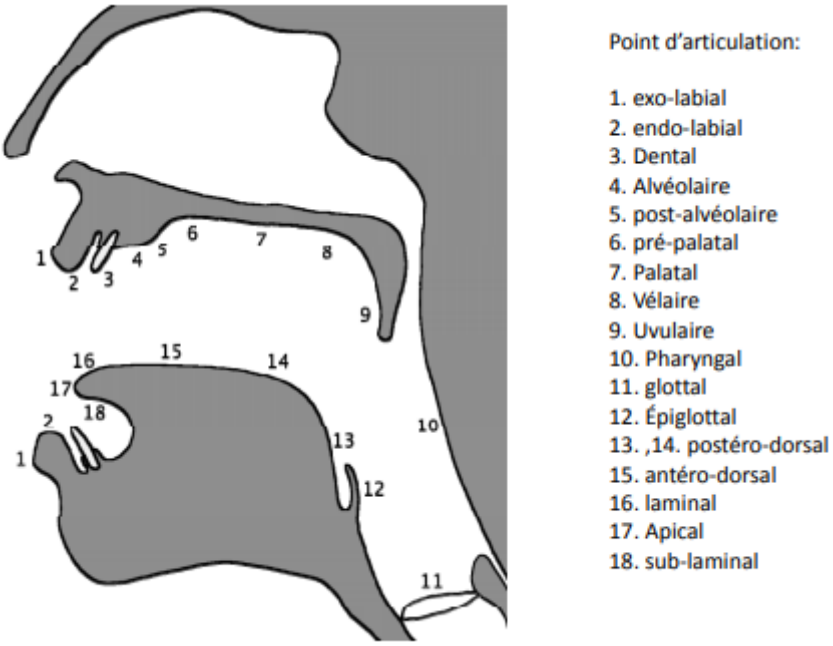


FIGURE 1.5 – Classification des consonnes.[2]

1.4 La synthèse de la parole

Cette partie décrit le principe de la production automatique de la parole aussi appelée souvent synthèse de la parole. Actuellement, de nombreuses applications exploitent des techniques de synthèse de parole comme par exemple l'aide à la navigation (GPS), les serveurs vocaux téléphoniques, la diffusion de messages dans les lieux publics ainsi que dans les moyens du transport (les métros, les trains...), la robotique humanoïde, les assistants virtuels des smartphones (Siri d'Apple, Cortana de Microsoft...), l'industrie du divertissement (jeu vidéo...) ou encore l'assistance aux personnes atteintes de déficiences visuelles (via par exemple les boites de messagerie vocales ou la lecture automatique des pages web et des messages...). La plupart de ces applications ont recours à des techniques de synthèse vocale à partir d'un texte.

2. Source : http://www.claudegabriel.be/

1.4.1 Les principes de la synthèse de la parole

Le processus de synthèse de parole consiste à générer automatiquement un signal de parole à partir d'un texte. Ce processus est fréquemment appelé TTS (*Text To Speech*). L'entrée d'un tel système est un texte écrit et sa sortie est un signal de parole. Le mécanisme de synthèse est divisé en deux grandes étapes (figure 1.6).

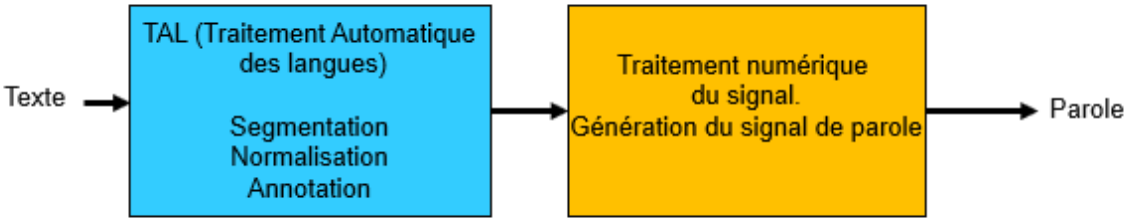


FIGURE 1.6 – Un aperçu d'un système TTS

La première étape consiste en une application des principes du traitement automatique de la langue naturelle sur le texte, à savoir la segmentation du texte en différents niveaux (phrases, mots, syllabes et phonèmes) (Taylor, 2009). En effet, le texte peut contenir différents types d'informations transcrites dont la procédure de synthèse nécessite la connaissance. Par exemple, le passage des graphèmes aux phonèmes, l'information d'accentuation qui concerne les syllabes. De même, le texte doit être normalisé (il s'agit d'une transformation du texte pour que les informations qu'il contient soient présentées sous une forme canonique pour une application en aval (Sproat *et al.*, 2001)).

La deuxième étape génère le signal de parole en se basant sur les informations obtenues par la première étape de traitement du texte que ce soit par une concaténation de segments de parole (dans le cas d'une approche de synthèse par concaténation) ou en utilisant des modèles paramétriques.

1.4.2 Méthodes de synthèse de la parole

Différentes approches de synthèse ont été développées pour réaliser la synthèse de la parole à partir du texte. Cette section décrit les approches de synthèses les plus populaires.

FIGURE 1.7 – Modélisation 3D du conduit vocal.[3]

Synthèse articulatoire

Les premières approches de synthèse vocale étaient basées sur une imitation du processus physique de la production de la parole. (Fant, 2012) définit la parole comme étant "la réponse du conduit vocal à une ou plusieurs sources des sons". Ainsi conformément à la modélisation source-filtre, l'approche de synthèse articulatoire est basée essentiellement sur une modélisation du conduit vocal (Engwall, 1999) (exemple figure 1.7).

(Carlson, 1995) a indiqué que la synthèse articulatoire de parole modélise le processus de production de la parole naturelle aussi précisément que possible. Ceci est accompli en créant un modèle synthétique de physiologie humaine. La modélisation géométrique du conduit vocal permet de simuler les mouvements des articulateurs (Wu et Hsieh, 2000). Ceci fournit un moyen pour tirer profit des propriétés du mécanisme de la production de la parole et de la phonétique. La précision d'un tel modèle est étroitement liée à l'acquisition des données et les mesures faites sur le conduit vocal. Pour ce faire, il existe différentes méthodes qui dépendent de l'application du synthétiseur et de certains facteurs de sécurité et de précision.

Méthodes statiques : Ces sont des méthodes qui consistent en des mesures instantanées du conduit vocal. La principale caractéristique de ces méthodes est que seuls des

3. Source : (Engwall, 1999)

échantillons isolés d'articulation peuvent être obtenus, mais elles sont incapables de représenter le mouvement. En outre, un problème commun pour l'acquisition des données est généralement lié à une articulation prolongée, ce qui peut entraîner des résultats non naturels. Cela peut être causé par la fatigue du participant, mais aussi par son anticipation à prolonger l'articulation et d'autres facteurs liés à la méthode. Très souvent, les articulations peuvent être vérifiées indirectement en comparant le son produit lors de l'acquisition de données avec le son produit dans des conditions plus naturelles. Pendant le traitement des données, les artefacts peuvent être supprimés en fixant le système de coordonnées sur des os plutôt que sur la position du sujet dans l'appareil de mesure.

Méthodes dynamiques : Ce sont des méthodes capables de détecter le mouvement articulatoire. Cette capacité a un coût en résolution spatiale ou nécessite de limiter la méthode à deux dimensions ou à un ensemble de points. Certaines considérations relatives aux méthodes statiques sont également valables pour les méthodes dynamiques. En particulier, la fatigue et les mouvements indésirables du sujet pendant l'acquisition sont des problèmes universels pour l'acquisition de données. En revanche, certaines sources d'erreur sont plus courantes avec la méthode dynamique qu'avec les méthodes statiques. L'une des plus importantes est l'articulation non naturelle résultant d'un équipement, qui doit être placé dans la cavité buccale du sujet.

Le modèle géométrique du conduit vocal nécessite la modélisation de trois parties, à savoir la géométrie de base, les paramètres de mouvement et le mécanisme de génération des mouvements :

— Modèle géométrique du conduit vocal.

— Modèle des paramètres du conduit vocal.

— Modèle de mouvement.

Synthèse par règles

L'approche de synthèse par règles est apparue parmi les premières techniques de synthèse de parole à partir d'un texte. Cette méthode exige la compréhension des mécanismes de perception et de production de parole. La technique de synthèse par formants est la plus utilisée des techniques de synthèse par règles.

Les formants sont les zones fréquentielles d'enveloppe maximale, ils correspondent aux fréquences de résonance de la fonction de transfert du conduit vocal. L'objectif de cette technique est de créer un signal de parole en se basant sur des caractéristiques des formants (amplitudes, fréquences centrales, largeurs de bande) et sur des règles d'évolution des formants entre les phonèmes (Klatt, 1980).

Cette technique est basée sur des règles déduites à la suite de l'observation des spectrogrammes de parole naturelle. Ces règles tracent la prononciation des phonèmes et l'évolution temporelle des formants ce qui permet de générer un spectre de signal de parole. Le processus de synthèse de la parole est décrit par une modélisation source filtre. Cette approche présente l'avantage de l'utilisation d'une taille de mémoire réduite , cependant la qualité de la parole générée ainsi que son intelligibilité sont dégradées par rapport à la parole naturelle.

Synthèse par concaténation

L'approche de synthèse par concaténation ne fait appel à aucune modélisation du conduit vocal. Un ensemble d'unités de parole pré-enregistrées est utilisé, le processus de synthèse est réalisé par une mise bout à bout des unités dans le bon ordre (Schroeter, 2004). Le choix des unités et la qualité de la parole enregistrée influent sur la qualité de la parole synthétisée (figure 1.8).

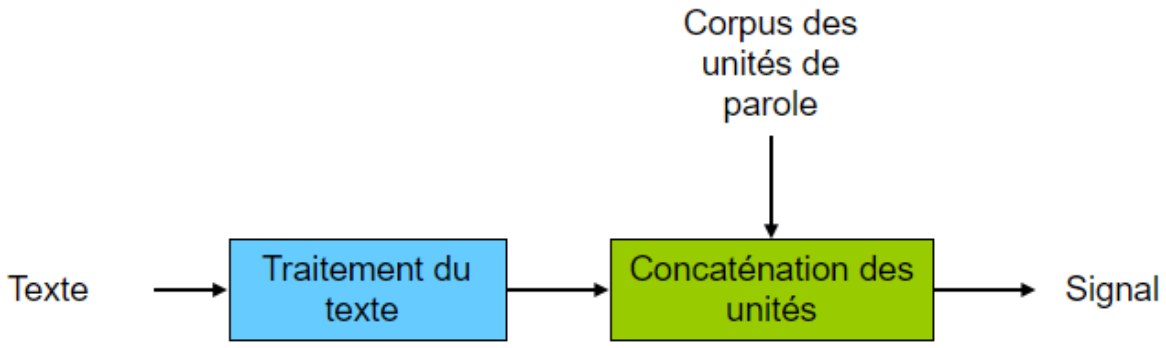


FIGURE 1.8 – Principe de synthèse par concaténation

Différentes tailles d'unités de parole peuvent être utilisées. Le choix le plus fréquent correspond aux diphones. Un diphone s'étend du milieu d'un phonème (sa partie stable)

au milieu du phonème suivant (Moulines et Charpentier, 1990). La concaténation est réalisée aux frontières des diphones dans des zones stables contrairement aux frontières des phonèmes qui sont des zones instables du fait des phénomènes de coarticulation.

PSOLA et MBROLA

Pendant les années 80, un algorithme de synthèse basé sur la modification de la fréquence fondamentale et de la durée du signal de parole sans toucher à l'identité des segments du signal appelé PSOLA (Pitch Synchronous Overlap Add) a été développé (Moulines *et al.*, 1990). Les modifications sont réalisées sans utiliser la modélisation source-filtre de la parole. En pratique la synthèse de parole utilise cet algorithme comme suit :

1. Isolation des périodes de la fréquence fondamentale au niveau du signal original.
2. Faire les modifications nécessaires.
3. Générer le signal final.

Il existe plusieurs variantes de cet algorithme, la plus utilisée est TD-PSOLA (Time-Domain Pitch-Synchronous Overlap-and-Add) (Moulines *et al.*, 1990). Elle est basée sur une synchronisation du pitch, ce qui veut dire qu'il y a une seule fenêtre d'analyse par période de pitch. La qualité de la parole synthétisée par TD-PSOLA est affectée par la localisation des périodes de pitch ; à la moindre erreur de localisation, la qualité sera dégradée.

Basé sur l'algorithme TD-PSOLA, le synthétiseur MBROLA (Multi-Band Resynthesis OverLap Add) a été conçu (Dutoit *et al.*, 1996). Le but principal était de trouver une solution au problème de localisation des périodes de pitch. MBROLA a recours à une technique de synthèse basée sur une modélisation harmonique/bruit ; il n'est pas nécessaire que ces positions soient cohérentes d'une trame à une autre. Pendant l'analyse, les trames sont retrouvées comme avant, mais lors de la synthèse, les phases sont ajustées de sorte que chaque trame de la base de données aura la phase correspondante. Cette étape permet d'ajuster efficacement toutes les périodes de pitch de manière à se trouver dans les mêmes positions relatives dans les trames.

Synthèse par sélection d'unités

À la suite des évolutions des mémoires de calculateurs, les méthodes de concaténation ont évolué. Au début, il y avait un unique exemplaire de chaque unité de parole (diphone).

Par la suite, une nouvelle version de la synthèse par concaténation appelée synthèse par sélection d'unités, permet d'avoir plusieurs exemplaires de chaque unité de parole dans le corpus enregistrées dans différents contextes phonétiques et prosodiques (Hunt et Black, 1996). Le processus de synthèse est décrit dans la figure 1.9.

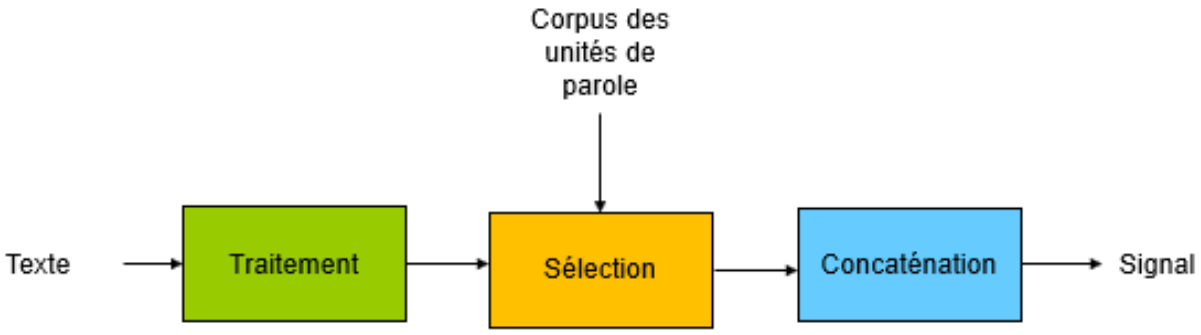


FIGURE 1.9 – Principe de synthèse par sélection d'unités

Pendant la phase de sélection, la meilleure séquence des unités est sélectionnée parmi les unités candidates du corpus. Ceci en se basant sur deux critères qui sont un coût de cible pour mesurer la similarité entre les caractéristiques des unités sélectionnées et les caractéristiques désirées, et un coût de concaténation pour mesurer la qualité de la concaténation. La synthèse par sélection d'unités a permis d'obtenir un saut qualitatif de la parole générée comparée à celle obtenue avec une synthèse par concaténation de diphones ou par règles.

Synthèse paramétrique

Cette approche de synthèse de parole est basée sur l'utilisation de modèles paramétriques. Le signal de parole est décrit par un ensemble de paramètres acoustiques extraits à des intervalles de temps réguliers. Ces paramètres contiennent essentiellement les paramètres acoustiques (les coefficients Mel-cepstraux et leurs dérivées temporelles, la fréquence fondamentale et la durée des phonèmes). Dans l'approche paramétrique statistique, ces paramètres sont décrits par des statistiques (les moyennes, les variances, les fonctions de densité de probabilité) pour représenter la distribution de ces paramètres. Un principe général de la synthèse par approche paramétrique est décrit dans la figure 1.10.

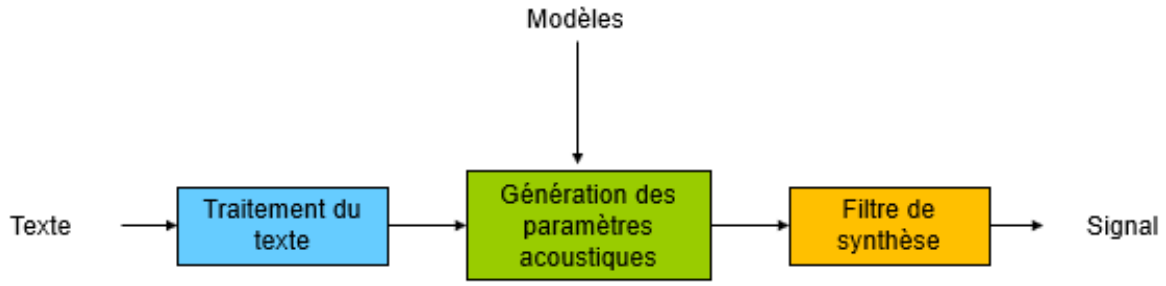


FIGURE 1.10 – Principe de synthèse par approche paramétrique

Le processus commence par un traitement du texte écrit (étape commune dans toutes les approches de TTS). L'étape suivante consiste en une prédiction des paramètres acoustiques qui seront traités par un synthétiseur (filtre de synthèse ou un vocodeur) afin de générer le signal de parole correspondant au texte à synthétiser.

Une première version de cette approche était basée sur l'utilisation des HMM (Black *et al.*, 2007). Le système résultant présente les avantages de l'utilisation d'une quantité de mémoire réduite par rapport à l'approche de synthèse par concaténation, ainsi que la possibilité de changer les caractéristiques de la voix générée (Yamagishi *et al.*, 2004). Depuis quelques années, les HMM ont été remplacés par des réseaux de neurones (DNN) dans l'approche de synthèse paramétrique (Zen et Senior, 2014). L'introduction de cette architecture (DNN) a permis d'améliorer la qualité de la parole générée. Ces approches sont détaillées dans le chapitre suivant.

1.5 Méthodes d'évaluation de la synthèse de la parole

Les systèmes de synthèse de parole sont conçus pour des fins spécifiques, ainsi, il est important de savoir si le système arrive à accomplir la tâche exigée. C'est l'étape d'évaluation du système de synthèse.

1.5.1 Principe d'évaluation

À ce niveau, les questions suivantes doivent être posées afin de définir les critères et les principes d'évaluation.

Pourquoi évaluer : Le but principal est d'améliorer les performances du système de synthèse conçu, ainsi l'évaluation permet de diagnostiquer ce qui a été réalisé afin de continuer le développement du système. De même, le système de synthèse peut être comparé à d'autres systèmes antérieurs pour le situer par rapport à l'état de l'art (Taylor, 2009), (Bennett, 2005).

Quand évaluer : La réponse à cette question dépend de l'application pour laquelle le système de synthèse a été conçu. Il est toutefois possible que l'évaluation ait lieu à différentes étapes du processus de synthèse : juger les performances d'un composant particulier du système, par exemple l'étape de traitement du texte ou bien évaluer le système de synthèse entier (Chevelu *et al.*, 2015).

Que faut-il évaluer : A ce niveau, ces sont les signaux générés qui sont évalués, en tenant compte de certaines aspects qui caractérisent la parole, telles que l'aspect naturel et l'intelligibilité de la parole synthétisée.

Comment évaluer : Cela consiste en une élaboration du plan des tests, les questions à poser aux participants ainsi que des mesures objectives qui ne font pas appel à l'intervention des auditeurs et sont basés sur des algorithmes qui mesurent par exemple la distance entre un signal de parole naturelle et son homologue synthétisé automatiquement par le système de synthèse à évaluer.

Comment analyser les résultats de l'évaluation : Les résultats récupérés en fin des tests (objectifs ou subjectifs) sont analysés ceci après avoir représenté les résultats sous une forme qui permettent de les analyser.

1.5.2 Critères à évaluer

L'accent est mis sur les signaux synthétisés plutôt que sur le système de synthèse. (Wester *et al.*, 2015) juge important la définition des critères à évaluer.

Qualité globale : Ce facteur fait référence à la qualité générale du signal généré par le système de synthèse. Dans certains cas, ce critère indique si l'auditeur doit faire un effort pour écouter le signal ou bien si c'était facile.

L'intelligibilité : Le message contenu dans les signal de parole synthétisé doit être compréhensibles. Les participants sont appelés à transcrire ce qu'ils viennent d'écouter (Sydeserff *et al.*, 1992).

L'aspect naturel : On cherche à évaluer la similarité entre la parole synthétisée et la parole naturelle prononcée par un être humain (Benoīt *et al.*, 1996).

Deux formes distinctes de tests peuvent être utilisées. Une première famille des tests objectifs qui sont essentiellement basés sur des procédures automatiques et des algorithmes de calcul des différences par rapport à des signaux de référence (en utilisant un calcul des distances par exemple). La seconde famille concerne les tests subjectifs. Pendant les expériences, les auditeurs interviennent et évaluent certains critères qui ne peuvent pas être jugés en utilisant les mesures objectives.

1.5.3 Évaluation objective

Cette méthode s'intéresse à la comparaison des propriétés acoustiques du signal synthétisé avec celles du signal de parole référence. Ce type des tests exige que les signaux synthétisés et ceux de référence soient alignées soit en fonction du contenu soit en fonction des durées. Le calcul des différences se fera ensuite à l'échelle des trames afin d'obtenir la distance totale entre les deux signaux. Les mesures objectives concernent essentiellement les propriétés du signal ; les coefficients spectraux, la prosodie, le voisement et la durée.

MCD (Mel Cepstral Distorsion) : La représentation spectrale permet de calculer la distorsion entre les deux signaux sur chaque trame d'alignement. La distorsion totale est obtenue en faisant la somme des distorsions calculées sur l'ensemble des trames. Plus la distorsion est grande, plus le signal généré est différent du signal naturel (Kubichek, 1993).

La prosodie (F0) : L'évaluation de la fréquence fondamentale peut se faire en calculant l'erreur quadratique moyenne RMSE (Root Mean Sqaure Error) (Chai et Draxler, 2014) sur l'ensemble des trames, ou en mesurant la corrélation entre les contours de F0 correspondants respectivement au signal naturel et au signal synthétisé.

La durée : Il est de même possible d'évaluer les durées des unités de parole synthétisées. Une première possibilité est le calcul de l'erreur quadratique moyenne entre la durée prédite et la durée de référence. Il est aussi possible d'évaluer la prédiction en calculant le rapport entre la durée prédite et la durée naturelle.

Voisement : Il est aussi possible d'évaluer la prédiction du voisement, ceci en déterminant les erreurs de voisement : prédiction d'une trame voisée alors qu'elle doit être non

voisée et vice versa. L'évaluation de la prédiction du voisement sert d'indication sur la modélisation de la prosodie.

Apériodicité : Ce paramètre acoustique est spécifique à l'utilisation de certains vocodeurs. Généralement, l'apériodicité est introduite afin de maintenir l'apport de la partie périodique par rapport à celle non périodique (bruit) du signal de parole avant l'application du filtre (Kawahara *et al.*, 2001). L'apériodicité peut être évaluée en calculant l'erreur quadratique moyenne entre les valeurs extraites du signal naturel et celles extraites du signal synthétisé.

Évaluations de la qualité du signal : Les algorithmes PESQ (Perceptual Evaluation of Speech Quality) (Rix *et al.*, 2001) et POLQA (Perceptual Objective Listening Quality Analysis). L'algorithme PESQ fonctionne comme suit : un modèle perceptuel, réalise un alignement temporel entre le signal original (naturel) et le signal synthétisé. Les mesures indiquant les différences entre les deux signaux sont calculées. Enfin, un modèle cognitif simule des tests d'écoute et modélise les différences. La sortie est un score compris entre -1 et 4,5.

POLQA fait appel à un modèle permettant de prédire des mesures objectives indiquant la qualité de la parole transmise, basé sur une analyse numérique du signal de parole. Les mesures objectives devraient se corréler avec les scores de qualité subjectifs obtenus lors de tests d'écoute subjectifs. Cette méthode utilise le signal de parole naturel comme référence lors de l'évaluation. (Beerends *et al.*, 2013) l'utilise pour évaluer la transmission du signal de parole via les canaux de communication en mesurant la distorsion du signal reçu par rapport au signal naturel transmis. (Cernak et Rusko, 2005) propose une adaptation de l'algorithme PESQ (Recommendation, 2001) pour l'évaluation des systèmes de synthèse de parole. Son étude porte sur l'évaluation de la corrélation entre les résultats des tests subjectifs et objectifs.

1. Acquisition des données, dans (Cernak et Rusko, 2005), des mots contenant les différents contextes possibles ont été enregistrés par deux locuteurs. Ces enregistrements serviront des références pour l'évaluation plus tard.
2. Différents systèmes de synthèse de parole ont été utilisés afin de générer automatiquement des signaux de parole correspondants aux textes des signaux de référence.
3. Une évaluation subjective (tests d'écoute impliquant des participants) des signaux générés.

4. L'étape suivante consiste en une évaluation objective des signaux générés (les mots) en utilisant l'algorithme PESQ et les signaux de référence. Des scores objectifs sont obtenus.
5. Finalement, la corrélation entre les scores obtenus lors des évaluations objectives et subjectives est calculée.

1.5.4 Évaluation subjective

Les différentes mesures objectives citées précédemment ne permettent pas d'évaluer certains critères des signaux de parole générés, tel que l'aspect naturel ou l'intelligibilité. Ainsi, il est important de conduire des tests perceptifs faisant intervenir la participation d'auditeurs. Afin de garantir le bon déroulement des tests, il faut que la tâche des auditeurs participants soit claire et bien définie. La difficulté des tests dépend du type de la tâche :

- ***"Simple"*** : évaluer un signal en lui attribuant un score (sa qualité globale), ou comparer un signal synthétisé à un signal naturel (évaluer son aspect naturel), ou encore évaluer son intelligibilité en demandant à l'auditeur d'écrire ce qu'il vient d'entendre.
- ***"Compliquée"*** : Dans le cas d'évaluation d'un facteur particulier, les participants doivent se concentrer sur un aspect particulier. Ceci exige que les auditeurs soient des experts du domaine ou qu'ils doivent passer par une étape d'apprentissage.

Parmi les méthodes d'évaluation subjectives les plus utilisées :

— ***Les tests MOS (Mean Opinion Score)*** (ITU, 1996) : Il s'agit d'une évaluation absolue qui fait partie de la famille de classement absolu des catégories ACR (Absolute Category Rating) ; à chaque fois l'auditeur fait l'écoute d'un seul signal pour l'évaluer. Une échelle de jugement de catégorie à cinq points est proposée pour évaluer un certain critère de la parole :

  — 5 Excellente

  — 4 Bonne

  — 3 Passable

  — 2 Médiocre

  — 1 Mauvaise

Avant de commencer le test, des instructions sont présentées aux participants afin de s'assurer du bon déroulement du test. Une étape préliminaire est indispensable afin que les participants s'adaptent aux conditions du test : il s'agit d'évaluer quelques exemples des échantillons. Ceci permet aux participants de mieux comprendre le but du test. A la fin de l'évaluation, des scores de type MOS sont obtenus. Il est à noter que les scores associés à la phase préliminaire ne doivent pas être pris en compte lors de l'analyse des résultats. Les tests MOS peuvent être utilisés afin de juger la qualité globale et l'aspect naturel des signaux générés par le système de synthèse de parole.

— ***Les tests DMOS (Differential Mean Opinion Score)*** (ITU, 1996) : Il s'agit d'une méthode d'évaluation de la dégradation appartenant à la famille de classement de dégradation des catégories DCR (Degradation Category Rating). C'est une version modifiée de la méthode ACR, qui offre une meilleure distinction des signaux de bonne qualité (ITU, 1996). Un signal de référence de qualité (exemple : le signal naturel) est à introduire avant chaque signal à évaluer. Contrairement à la méthode précédente, les signaux sont présentés par paire pour chaque participant : A-B, A est le signal de référence et B est celui à évaluer. Les participants sont ensuite appelés à évaluer la qualité du deuxième signal selon une échelle de dégradation à cinq points :

  — 5 Dégradation inaudible

  — 4 Dégradation audible mais non gênante

  — 3 Dégradation un peu gênante

  — 2 Dégradation gênante

  — 1 Dégradation très gênante

  Des scores de type DMOS sont obtenus à la fin du test, ces scores sont analysés pour évaluer la dégradation que présente le signal de parole généré par rapport au signal naturel.

— ***Tests de Comparaison :*** Il s'agit d'un test de préférence faisant partie de la famille d'évaluation par comparaison de catégories CCR (Comparaison Category Rating) (ITU, 1996). Le principe de cette méthode est similaire à celui de la méthode précédente (DMOS) ; les signaux sont présentés par paire à chaque participant mais sans imposer un ordre particulier. Cependant, le but du test est la comparaison des variantes d'un même système de synthèse ou de deux systèmes différents sans faire recours à une référence. Les auditeurs pointent vers le signal préféré en évaluant la

qualité du deuxième signal écouté par rapport au premier selon l'échelle de préférence suivante :

- 7 Bien meilleure
- 6 Meilleure
- 5 Un peu meilleure
- 4 À peu près la même
- 3 Un peu plus mauvaise
- 2 Plus mauvaise
- 1 Beaucoup plus mauvaise

— ***Les tests MUSHRA*** (MUltiple phrase with Hidden Reference and Anchor) : C'est une méthode d'évaluation subjective à plusieurs étapes qui permet d'évaluer l'aspect naturel et la qualité des signaux de parole générés (Mason, 2002). Pendant le déroulement de ce test, les signaux de références sont inconnus et c'est à l'auditeur de choisir quel signal doit être la référence pour passer aux étapes suivantes de l'évaluation (Schoeffler *et al.*, 2015). Le principal avantage par rapport au test MOS est que MUSHRA nécessite moins de participants pour obtenir des résultats statistiquement significatifs. En outre, une échelle de 0 à 100 est utilisée par MUSHRA ce qui permet d'évaluer de très petites différences avec plus de précision.

1.6 Conclusion

Ce chapitre a décrit l'état de l'art de la synthèse de parole. Tout d'abord, une description physiologique du mécanisme de production de parole ainsi que les organes participants a été faite. Ensuite, le principe de la synthèse de parole à partir d'un texte a été décrit, en présentant les approches les plus utilisées. Enfin, la dernière partie a été consacrée à la présentation des différentes méthodes d'évaluation des systèmes de synthèse de parole.

Chapitre 2

Synthèse par approche paramétrique

Sommaire

2.1 Introduction

Après la description de différentes méthodes de synthèse de la parole à partir du texte dans le chapitre précédent, l'intérêt est particulièrement porté sur l'approche paramétrique dans ce chapitre. Tout d'abord, une description de l'aspect général ainsi que ses particularités telles que la paramétrisation du signal de parole, l'utilisation de vocodeurs pour l'analyse des signaux ainsi que les descripteurs contextuels sont présentés. L'approche de synthèse paramétrique peut faire appel à l'utilisation des HMM (Hidden Markov Model) ou des DNN (Deep Neurel Network). Les deux méthodes sont décrites en détail dans les parties qui suivent (principe de fonctionnement et outils utilisés pour l'implémentation de ces approches).

2.2 Synthèse par approche paramétrique

La synthèse par approche paramétrique repose sur une description du signal de parole par un ensemble de paramètres. Ces derniers sont modélisés par des modèles paramétriques qui serviront ensuite à la génération de la forme d'onde du signal de parole correspondant au texte à synthétiser. Cette approche nécessite une description de chaque segment du texte à synthétiser par un ensemble d'informations appelées descripteurs contextuels afin de distinguer les différents contextes dans lesquels ce segment peut exister.

2.2.1 Aspect général

La figure 2.1 présente les blocs principaux de l'approche de synthèse paramétrique de la parole. L'entrée est un texte écrit. Le processus nécessite une paramétrisation des données d'apprentissage qui seront ensuite représentées par des modèles paramétriques ou statistiques. Une phase de traitement du texte à synthétiser a pour but la qualification de chaque son du texte par un ensemble d'informations affectant la prononciation de celui-ci.

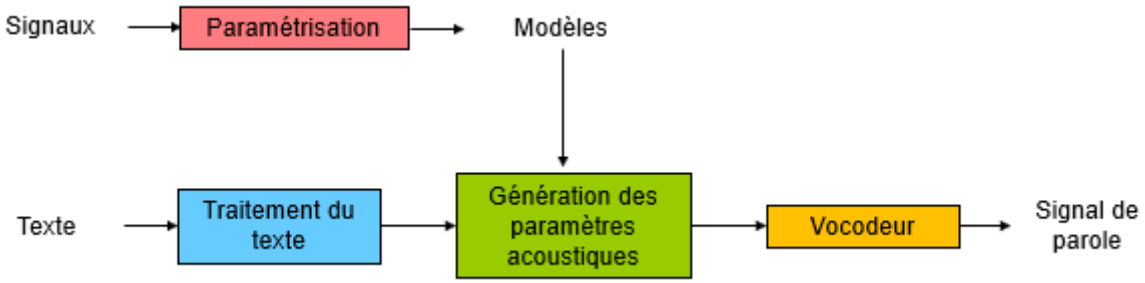


FIGURE 2.1 – Principe de la synthèse paramétrique

En se basant sur les modèles appris et les descripteurs contextuels, les paramètres acoustiques du signal sont générés puis fournis à un vocodeur pour générer le signal de parole correspondant au texte de l'entrée.

2.2.2 Paramétrisation du signal de parole

Les paramètres à choisir pour décrire le signal de parole doivent permettre la reconstruction du signal de parole. La représentation fréquentielle du signal (se basant sur la modélisation source/filtre de la production de la parole) permet d'extraire la fréquence fondamentale F0 ainsi que les réponses fréquentielles du conduit vocal formant ainsi l'enveloppe spectrale. En pratique, les vocodeurs permettent d'analyser les signaux de parole et de les représenter par les paramètres décrits ci-dessus.

2.2.3 Les vocodeurs

Le codage du signal permet de représenter le signal par un ensemble de paramètres. Un vocodeur permet d'analyser un signal par extraction de des paramètres acoustiques. Il est possible de générer ce même signal en utilisant les paramètres extraits en phase d'analyse (figure 2.2).

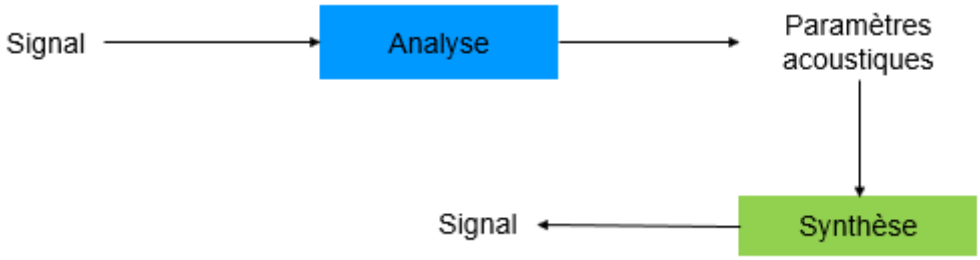


FIGURE 2.2 – Analyse/Synthèse

STRAIGHT

Le vocodeur STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) (Kawahara *et al.*, 1999) est largement utilisé dans l'approche paramétrique de synthèse de parole. STRAIGHT permet d'analyser le signal de parole en se basant sur la modélisation source/filtre afin d'extraire les paramètres qui permettront la reconstruction du signal. La taille de la fenêtre d'analyse est un compromis à considérer lors de l'estimation de l'enveloppe spectrale. Le but est d'obtenir une enveloppe spectrale dépourvue de tout information liée à la périodicité

D'après (Kawahara *et al.*, 1999), si la fenêtre est de taille réduite (comparable à la période fondamentale, T0=1/F0), le spectrogramme du signal présentera des fluctuations dans le domaine temporel. Quand la fenêtre (de petite taille) glisse tout au long du signal, la puissance subit une augmentation et une diminution successives à chaque période, ainsi une bonne résolution temporelle est obtenue. Dans le cas contraire, si la fenêtre d'analyse est large ce qui mène à une énergie quasiment constante, il n'y a pas de fluctuations dans le domaine temporel et une bonne résolution fréquentielle est obtenue.

STRAIGHT propose une fenêtre d'analyse à caractère adaptatif à la fréquence fondamentale. En variant la taille de la fenêtre, l'énergie reste quasiment constante. Cette méthode permet de minimiser l'interférence entre l'enveloppe spectrale et les harmoniques (Kawahara *et al.*, 1999). L'analyse de Fourier est ensuite appliquée suivie d'un lissage pour éliminer les harmoniques restantes.

Afin de tenir compte de l'a-périodicité dans certaines parties du signal de parole (sons non voisés), STRAIGHT introduit un troisième paramètre (en plus de la fréquence fondamentale et des paramètres spectraux). L'a-périodicité est estimée en calculant la différence entre l'enveloppe spectrale supérieure (les composantes périodiques du signal)

et l'enveloppe inférieure (les composantes a-périodiques). Ainsi l'analyse du signal par STRAIGHT extrait la fréquence fondamentale, les coefficients spectraux et les paramètres d'a-périodicité. Les dimensions des paramètres spectraux sont très élevées. Pour les réduire, un filtre de fonction continue est utilisée (MLSA par exemple) pour estimer les coefficients spectraux dans l'espace. Le nombre de bandes dépend du vocodeur, STRAIGHT utilise 5 bandes et WORLD utilise 25 bandes. Les paramètres d'apériodicité sont réduits en divisant le spectre en des bandes larges de fréquences et calculer la moyenne d'énergie à travers ces bandes à chaque trame.

WORLD

Le vocodeur WORLD (Morise *et al.*, 2016) a été conçu pour générer de la parole dans le but d'améliorer la qualité des signaux synthétisés. WORLD est basé sur trois algorithmes pour analyser le signal et un algorithme pour assurer la phase de synthèse (ou la reconstruction du signal à partir des paramètres extraits) (figure 2.3).

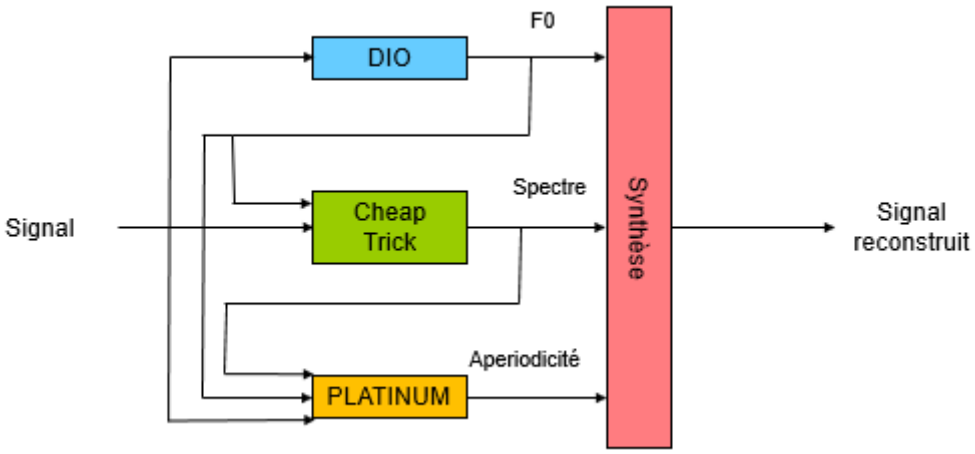


FIGURE 2.3 – Analyse et reconstruction avec WORLD

Pendant la phase d'analyse, la fréquence fondamentale est estimée en utilisant l'algorithme DIO ((Morise *et al.*, 2009), (Morise *et al.*, 2010)). L'estimation se fait en trois étapes :

1. Appliquer un filtre passe-bas en utilisant différentes fréquences de coupure. Si le signal obtenu est formé uniquement des composantes fondamentales, il forme une

sinusoïde de période T0 (période fondamentale).

2. Le deuxième étape est le calcul des candidats de F0 et leur fiabilité dans chaque signal filtré.
3. Le candidat de F0 qui obtient la fiabilité la plus élevée est sélectionné.

Pour estimer les paramètres du spectre, WORLD utilise la forme d'onde du signal et la fréquence fondamentale F0. En pratique, l'algorithme CheapTrick ((Morise, 2015a), (Morise, 2015b)) est utilisé pour estimer l'enveloppe spectrale. Il est basé sur une analyse de pitch synchrone en utilisant une fenêtre de Hanning de longueur 3*T0. La puissance du spectre est calculée, stabilisée puis lissée en utilisant une fenêtre rectangulaire de largeur $4\pi/3$*T0. WORLD utilise PLATINUM (Morise, 2012) pour l'extraction des paramètres d'apériodicité en se basant sur le signal de parole, F0 et l'enveloppe spectrale. PLATINUM applique une fenêtre de longueur 2*T0 au signal.

2.2.4 Les descripteurs contextuels

D'après (Taylor, 2009), la représentation acoustique d'un segment du texte (généralement un phonème), dans un certain contexte est étroitement liée non seulement à ses phonèmes voisins mais aussi d'autres facteurs. Ainsi, la spécification contextuelle du texte à synthétiser n'est pas limitée à une séquence des phonèmes formant les différentes phrases du texte. L'ensemble de descripteurs comprend tous les facteurs qui peuvent affecter la prononciation de chaque son du texte. Différents types d'information sont impliqués : des informations linguistiques, phonologiques et prosodiques. L'extraction de ces descripteurs se fait à plusieurs échelles : des informations selon la syllabe, le mot, l'expression ou la phrase auxquels appartient le segment du texte. (Tokuda *et al.*, 2002) a défini un ensemble standard de descripteurs contextuels (environ 50 descripteurs) qui regroupe les différents types d'information nécessaires à la description d'un segment de parole dans un contexte particulier (Annexe A) :

Descripteurs à l'échelle du phonème : La description d'un segment du texte à ce niveau est formée essentiellement par une séquence de cinq symboles (phonèmes ou silence) dont le symbole central de cette séquence révèle l'identité phonétique du segment courant, le reste de la séquence comporté les identités des deux phonèmes qui suivent le phonème central ainsi que celles des deux phonèmes qui le précèdent. La description à ce niveau comprend aussi la position du phonème courant par

rapport au début et à la fin de la syllabe à laquelle il appartient.

Descripteurs à l'échelle de la syllabe : À ce niveau, seules les syllabes courante, précédente et suivante sont concernées. Pour chaque syllabe, un descripteur est utilisé pour préciser sa longueur en nombre de phonèmes. Deux autres descripteurs sont utilisés pour donner une information sur la nature de l'accent de chaque syllabe : accent lexical ou accent tonique.

L'accent lexical ("stress") est une propriété inhérente des mots, la position de la syllabe qui la porte peut-être prédite selon des règles (Black *et al.*, 1998). Par ailleurs, l'accent tonique indique les phénomènes intonatifs (associés au pitch) (Taylor, 2000) au niveau de la syllabe. D'autres descripteurs sont utilisés afin de situer la syllabe courante par rapport au mot, à la phrase et à l'énoncé auxquels elle appartient. Des descripteurs complémentaires sont introduits afin de repérer la syllabe courante selon l'accentuation qu'elle porte et celles des syllabes voisines :

- Nombre de syllabes accentuées (accent lexical) qui précèdent et suivent la syllabe courante dans la phrase à laquelle elle appartient.
- Nombre de syllabes accentuées (accent tonique) qui précèdent et suivent la syllabe courante dans la phrase à laquelle elle appartient.
- Nombre de syllabes qui séparent la syllabe courante de la syllabe accentuée (accent lexical) qui la précède.
- Nombre de syllabes qui séparent la syllabe courante de la syllabe accentuée (accent tonique) qui la précède.
- Nombre de syllabes qui séparent la syllabe courante de la syllabe accentuée (accent lexical) qui la suit.
- Nombre de syllabes qui séparent la syllabe courante de la syllabe accentuée (accent tonique) qui la suit.

Finalement, un dernier descripteur identifie la voyelle de la syllabe courante.

Descripteurs à l'échelle du mot : À ce niveau, les premiers descripteurs sont utilisés pour indiquer le nombre de syllabes dans le mot courant, le mot précédent et le mot suivant ainsi que la position du mot courant dans la phrase courante. L'étiquette grammaticale du mot courant, précédent et suivant est prise en compte : le descripteur correspondant indique s'il s'agit d'un auxiliaire de temps, un auxiliaire modal, une

conjonction, un déterminant, un pronom personnel, un pronom interrogatif, une préposition ou une ponctuation.

Un descripteur a été introduit pour distinguer les mots signifiants ceci en faisant référence à son étiquette grammaticale qui permet de déterminer si un mot est grammatical ou non (Le Maguer *et al.*, 2013). À cette information, s'ajoutent d'autres descripteurs qui positionnent le mot courant par rapport aux mots signifiants qui le précèdent et le suivent dans la phrase courante.

Descripteurs à l'échelle de phrase : Les informations relatives à la phrase sont le nombre de syllabes et de mots dans les phrases courante, précédente et suivante. À ces informations s'ajoute un descripteur qui indique la prosodie de la fin de la phrase **ToBI** (Silverman *et al.*, 1992). ToBI est défini par un ensemble de conventions pour la transcription et l'annotation de la prosodie du langage (Beckman *et al.*, 2005).

L'information de la prosodie peut être définie par les phénomènes d'intonation ou les indices de rupture entre les mots. La première possibilité est basée sur la description de la courbe de la fréquence fondamentale (F0) par une séquence de symboles mélodiques qui indiquent le degré de la tonalité : **L** (tonalité basse), **H** (tonalité haute) et le symbole % pour indiquer le début ou la fin de l'énoncé). La seconde manière de décrire la prosodie de la fin de la phrase consiste à décrire les indices de rupture ou les frontières entre les mots (Port, 2008). Ceci par l'intermédiaire d'une échelle de 0 (frontière clitique) à 4 (fin de la phrase).

Descripteurs à l'échelle de l'énoncé : L'énoncé est décrit par le nombre de syllabes, des mots et des phrases qui le forment.

Une partie des descripteurs dépend de la langue. Ainsi, l'application de l'approche paramétrique de synthèse de parole à une langue nécessite la connaissance de ses particularités phonologiques, linguistiques et prosodiques. L'ensemble des descripteurs sera ensuite adapté selon les caractéristiques de la langue. Les modifications apportées peuvent être la suppression des descripteurs qui sont jugés inutiles pour la langue cible, l'ajout des nouveaux descripteurs afin de tenir compte des particularités de la langue cible ou l'ajout d'autres échelles de descriptions contextuelles des segments du texte.

2.3 Synthèse paramétrique statistique

La synthèse paramétrique de parole est basée sur l'utilisation de modèles pour représenter les paramètres du signal. Ces modèles peuvent être statistiques, basés sur l'utilisation des statistiques (la moyenne, la variance, la fonction de densité de probabilité) pour représenter les paramètres extraits du signal. Les HMM qui ont été d'abord utilisés pour la reconnaissance automatique de la parole (Young et Young, 1993), constituent la base d'une approche paramétrique de synthèse de parole (Black *et al.*, 2007).

Les HMM sont des modèles statistiques qui génèrent une séquence d'observations en utilisant des états cachés. Un exemple d'un modèle HMM à trois états est présenté dans la figure 2.4.

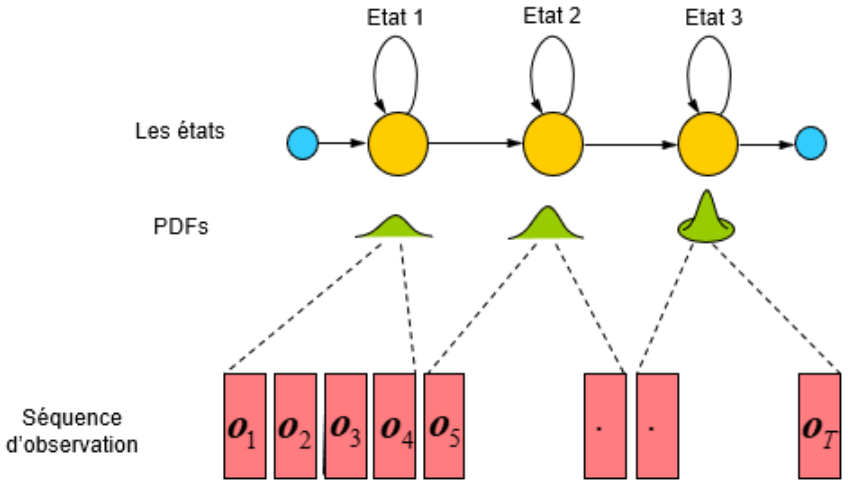


FIGURE 2.4 – Séquence d'observations HMM (Hidden Markov Model)

Dans un HMM, une fonction de densité de probabilité (PDF) est associée à chaque état. Elle décrit la distribution des paramètres observés correspondants à l'état. En synthèse de parole, chaque observation est un vecteur des paramètres acoustiques (les paramètres spectraux, F0,..). En raison de leur robustesse en reconnaissance automatique de la parole, les HMM ont aussi été utilisés dans la synthèse paramétrique de la parole ((Zen *et al.*, 2007), (Black *et al.*, 2007)). Les modèles HMM sont appris en utilisant des signaux de parole pour l'apprentissage. Les modèles HMM modélisent l'ensemble des paramètres décrits dans les sections précédents (les paramètres du spectre, la fréquence fondamentale, la durée). Des statistiques (moyennes et variances) sont utilisées pour décrire les distributions des

paramètres des données d'apprentissage. HTS (HMM-based Speech Synthesis System ou H Triple S) est un système de synthèse de parole par HMM et est basé sur l'utilisation de HTK.

(Zen *et al.*, 2007) a montré que HTS présente certains avantages par rapport aux autres systèmes de synthèse de parole principalement le fait d'être entraînable, et la possibilité de changer les caractéristiques de la voix générée. Ces avantages et d'autres seront détaillés dans la section 2.3.7. HTS a été adapté à plusieurs langues, telles que l'anglais (Tokuda *et al.*, 2002), le japonais (Zen, 2006), l'allemand (Krstulovic *et al.*, 2007) et le français (Le Maguer *et al.*, 2013). HTS est basé sur l'utilisation de modèles de phonèmes dépendants du contexte ; ceci exige la description de chaque unité de parole par un ensemble de descripteurs contextuels tels que ceux décrits précédemment. Cette partie décrit l'aspect général de la synthèse de parole par les HMM, ses particularités et les méthodes de modélisation introduites ainsi que ses avantages.

2.3.1 Aspect général

Le processus de HTS est divisé en deux phases : apprentissage et synthèse comme décrit dans la figure 2.5. L'apprentissage commence par extraire les paramètres acoustiques des signaux de parole du corpus d'apprentissage. Les paramètres sont extraits toutes les 5ms. Dans la plupart des cas, l'extraction est faite en utilisant le vocodeur STRAIGHT. L'ensemble des paramètres comporte les paramètres du spectre (les coefficients mel cepstraux et leur dérivées premières et secondes), les paramètres d'excitation (log F0 et ses dérivées) et les paramètres d'apériodicité (5 bandes). Les paramètres acoustiques extraits sont modélisés en utilisant des HMM dépendants du contexte. Chaque son correspondant aux signaux d'apprentissage est décrit par un ensemble de descripteurs contextuels. Ainsi une séquence de descripteurs est obtenue qui correspond à la séquence des phonèmes du texte. A la fin de l'apprentissage, trois modèles sont obtenus : un modèle pour la durée, un pour les paramètres du spectre et un pour la fréquence fondamentale.

La phase de synthèse commence par la conversion du texte à synthétiser en une séquence de descripteurs contextuels. La "phrase" HMM correspondante est construite par concaténation des HMM dépendants du contexte correspondant à la séquence des descripteurs contextuels. Les durées des états correspondants sont prédites de manière à maximiser leur probabilité. Les paramètres du spectre, et d'excitation (la fréquence fondamentale) sont générés à partir des HMM en utilisant l'algorithme de génération des

paramètres à partir des densités de probabilité. Enfin, le filtre de synthèse MLSA (Mel-Log Spectrum Approximation) génère le signal de parole en utilisant les paramètres acoustiques prédits.

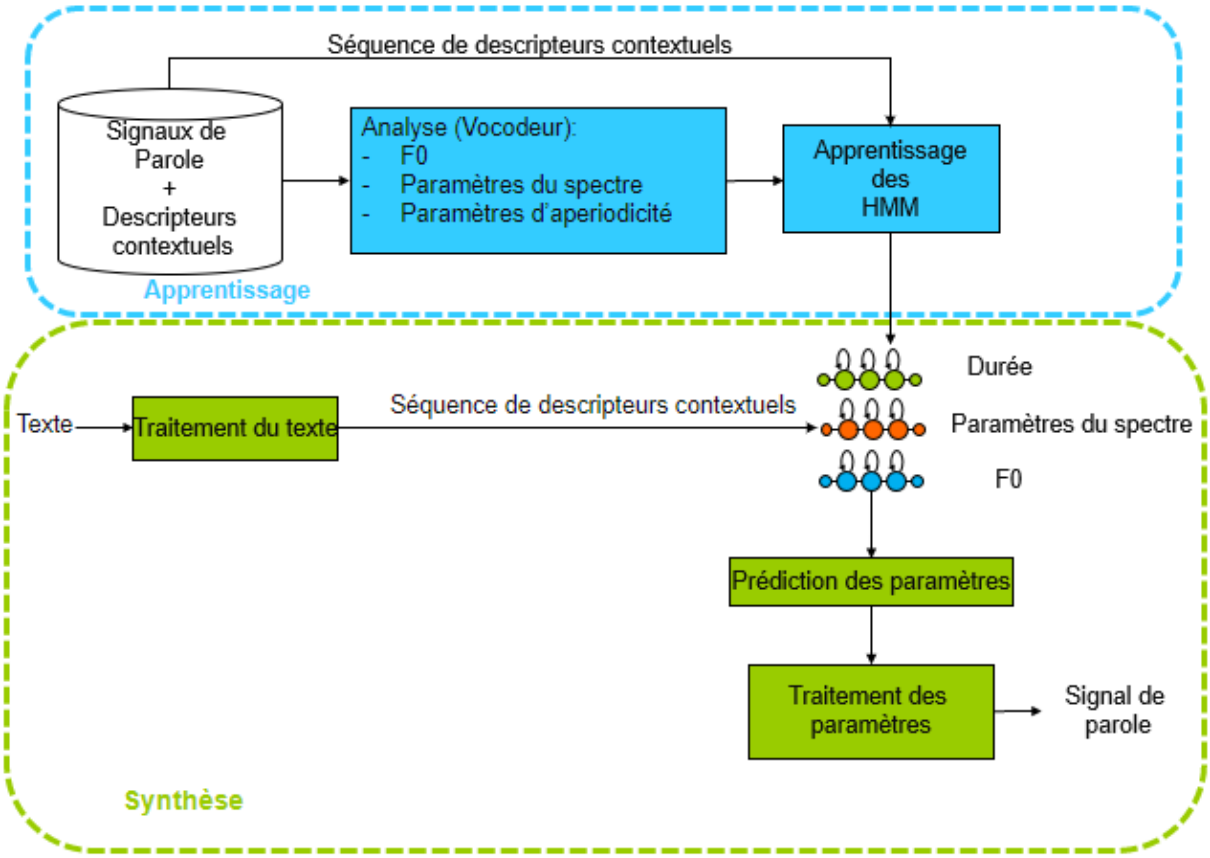


FIGURE 2.5 – Aspect général du système HTS

2.3.2 Modélisation dépendante du contexte

Les paramètres acoustiques d'un phonème sont étroitement liés à son contexte : principalement les phonèmes voisins. En pratique, ce n'est pas possible d'avoir des données d'apprentissage qui couvrent tous les contextes possibles dans lesquels un phonème peut apparaître. Ainsi il y a des contextes qui n'apparaissent qu'un nombre réduit de fois, et d'autres contextes qui n'apparaissent jamais dans les données d'apprentissage. Dans ces deux cas, il n'y aura pas suffisamment des données pour apprendre les modèles HMM associés à ces contextes, ce qui affectera la génération des paramètres acoustiques du

signal de parole à synthétiser. Une solution est de partager les paramètres entre densités similaires c'est-à-dire : entre contextes similaires.

Les densités similaires sont groupées et représentées par une seule densité partagée. La similarité des modèles est jugée selon les facteurs contextuels : des contextes sont similaires s'ils ont le même effet sur la prononciation du phonème courant. Les arbres de décision permettent de grouper les densités similaires selon les descripteurs contextuels (Jurafsky et James, 2000). Comme la fréquence fondamentale, le spectre et la durée sont affectés par des facteurs contextuels distincts, le groupement des densités pour chaque paramètre se fait séparément : il y aura un arbre de décision pour le spectre, un arbre pour la fréquence fondamentale et un arbre global de durée pour tous les états (figure 2.6).

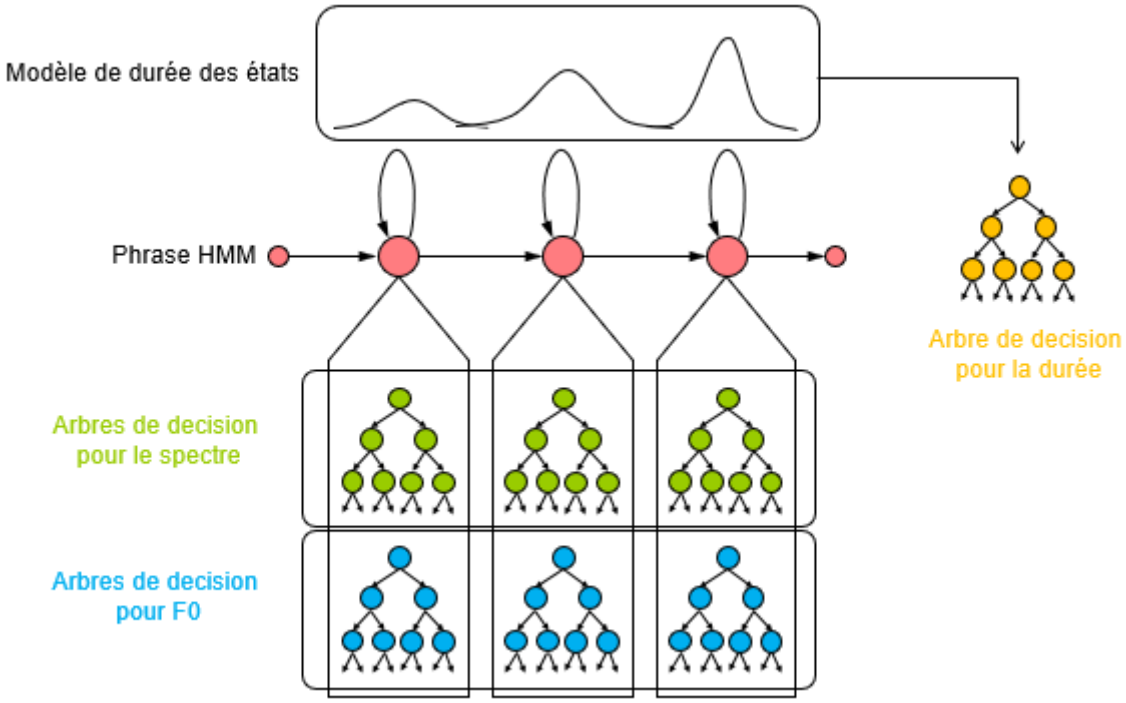


FIGURE 2.6 – Arbres des décisions

Un arbre de décision fonctionne de manière descendante. Chaque nœud correspond à un test sur un descripteur contextuel. A chaque descripteur est associé un ensemble de valeurs pour décrire l'information linguistique ou prosodique. Les paramètres acoustiques sont obtenus par descente dans l'arbre jusqu'à rencontrer une feuille. Chaque feuille contient une

distribution statistique. Pour arriver à une feuille, il faut valider les nœuds de l'arbre qui contiennent les caractéristiques linguistiques ou prosodiques. Les densités jugées similaires sont groupées. Toutes les distributions regroupées sur une feuille sont remplacées par une seule distribution qui sera alors partagée par plusieurs états des HMM.

2.3.3 Les arbres de décision

Cette section décrit le processus de la construction d'un arbre de décision. Le mécanisme est similaire pour tous les paramètres (durée, spectre, fréquence fondamentale) (Taylor, 2009).

1. Créer un groupe initial contenant toutes les instances d'un contexte particulier.
2. Définir un ensemble des questions sur les descripteurs contextuels qui sont associés aux sons.
3. Pour chaque question :
   (a) Former deux nouveaux groupes selon la réponse à la question (vrai ou faux).
   (b) Mesurer l'impureté des deux nouveaux groupes.
4. Identifier la question qui mène à une réduction de l'impureté.
5. Supprimer cette question de la liste des questions à parcourir.
6. Former deux nouveaux groupes basés sur cette question.
7. Répétez les étapes de 3 à 6 sur chaque nouveau groupement jusqu'à ce que les critères d'arrêt soient remplis.

Les critères d'arrêt impliquent généralement une vérification de profondeur de l'arbre et de la taille des groupes. HTS utilise un critère de type MDL (Minimum Description Length) (Shinoda et Watanabe, 2001) pour contrôler la construction des arbres de décision. Le critère MDL permet de déterminer une taille d'arbre, il permet d'obtenir un compromis entre la précision de la modélisation et la parcimonie de description du modèle. MDL permet d'obtenir le nombre de feuilles qui minimise la longueur de description associée à l'arbre. Pour construite l'arbre de décision de la durée, seuls les descripteurs à nature phonétique (pour les sons) et les descripteurs linguistiques (pour les pauses) sont considérés (figure 2.7).

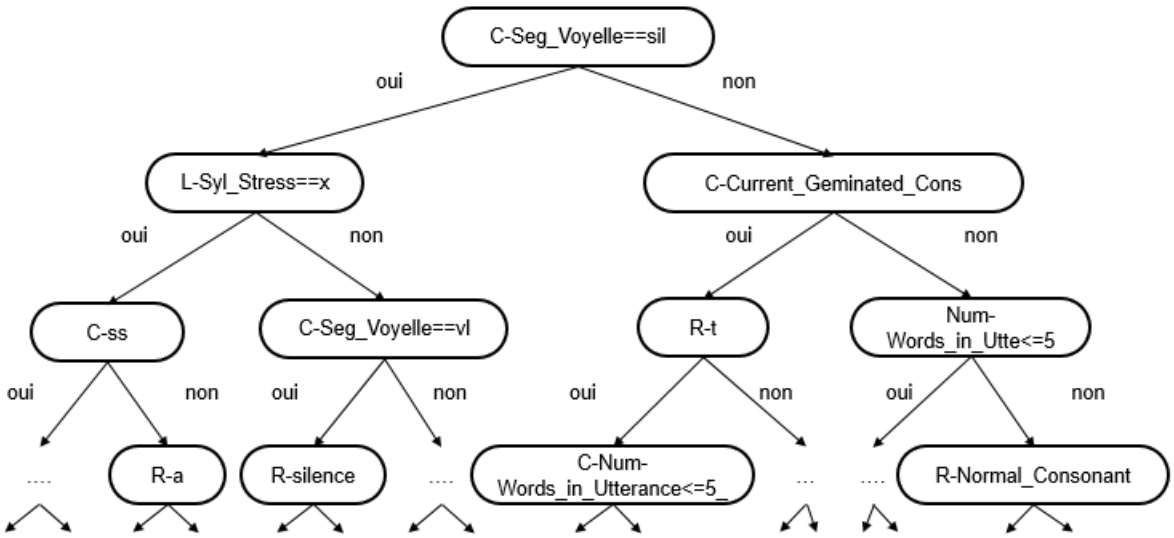


FIGURE 2.7 – Arbre de décision pour la durée des états (1er état).

Pour les arbres associés à la fréquence fondamentale F0, seuls les descripteurs à propos des attributs linguistiques sont pris en compte (figure 2.8).

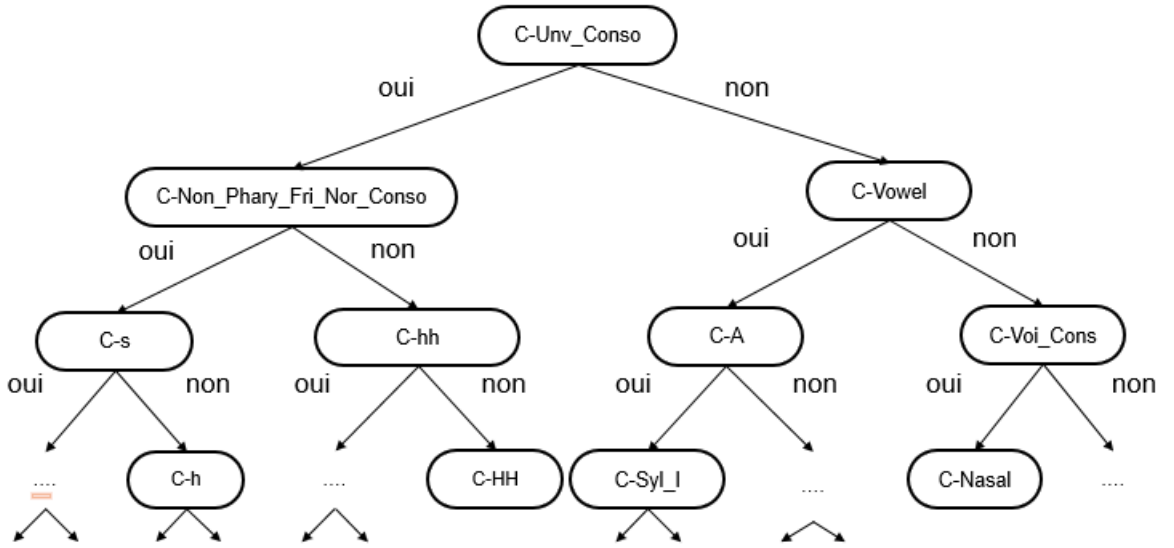


FIGURE 2.8 – Arbre de décision pour la fréquence fondamentale (1er état).

Finalement, pour les paramètres du spectre, seuls les descripteurs à propos des attributs phonétiques sont utilisés pour la construction de l'arbre de décision correspondante (figure 2.9).

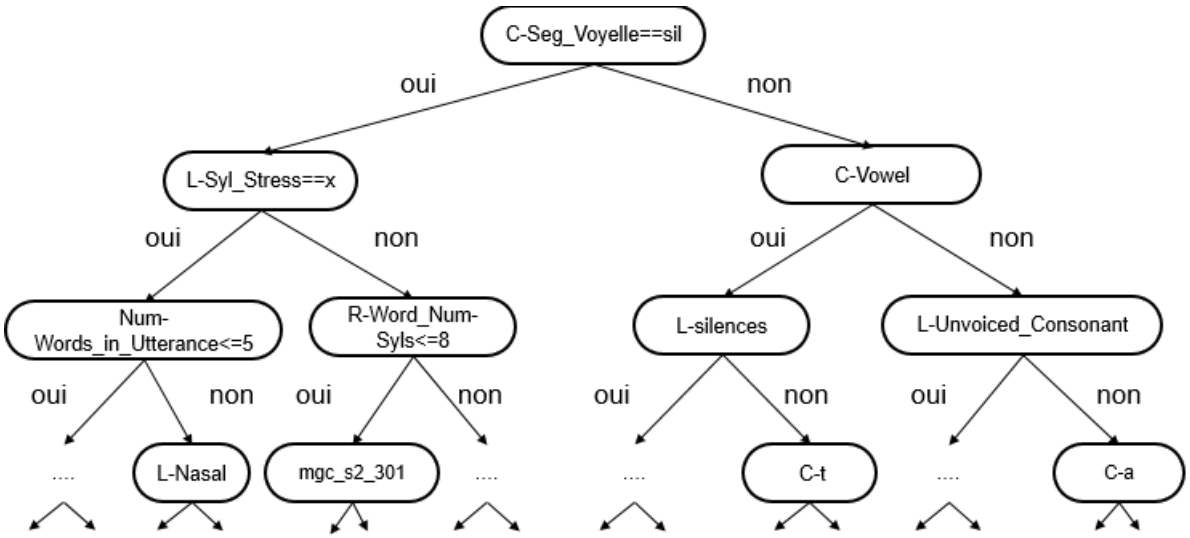


FIGURE 2.9 – Arbre de décision pour les paramètres du spectre (1er état).

2.3.4 Modélisation de la durée

Dans un modèle HMM standard, les probabilités de transition déterminent les caractéristiques de durée du modèle, c'est la durée de chaque état en nombre de trame (Taylor, 2009). La probabilité de durée des états est alors une exponentielle qui décroît au cours du temps ce qui attribue une durée d'une seule trame pour chaque état (Zen et Toda, 2005). Une solution pour remédier à ce problème, consiste à utiliser des modèles explicites pour la durée . (Yoshimura *et al.*, 1998) a proposé une solution basée sur l'utilisation des gaussiennes.

Les densités des durées des états sont estimées à partir des variables statistiques puis regroupés en utilisant un arbre de décision. Pendant la phase de synthèse, après construction de la phrase HMM, la durée de ses états est déterminée en maximisant leurs probabilités ensuite les paramètres acoustiques sont générés. Cependant, ce processus manque de consistance, car les paramètres d'apprentissage des HMM sont estimés sans modèle explicite de durée, les paramètres acoustiques sont générés à partir des HMM en

se basant sur un modèle explicite de durée. L'utilisation des modèles semi-Markoviens HSMM (Hidden Semi-Markov Models) (Levinson, 1986) vise à résoudre ce problème.

2.3.5 Modélisation des paramètres acoustiques

Pendant la phase de synthèse, les paramètres acoustiques sont générés à partir des modèles HMM (F0, paramètres du spectre) en utilisant l'algorithme de génération des paramètres MLPG (Maximum Likelihood Parameter Generation). Le principe de MLPG est utilisé pour générer une séquence d'observations qui sont les paramètres à introduire dans le vocodeur pour générer le signal de parole. Chaque état est décrit par une moyenne et une variance (figure 2.10).

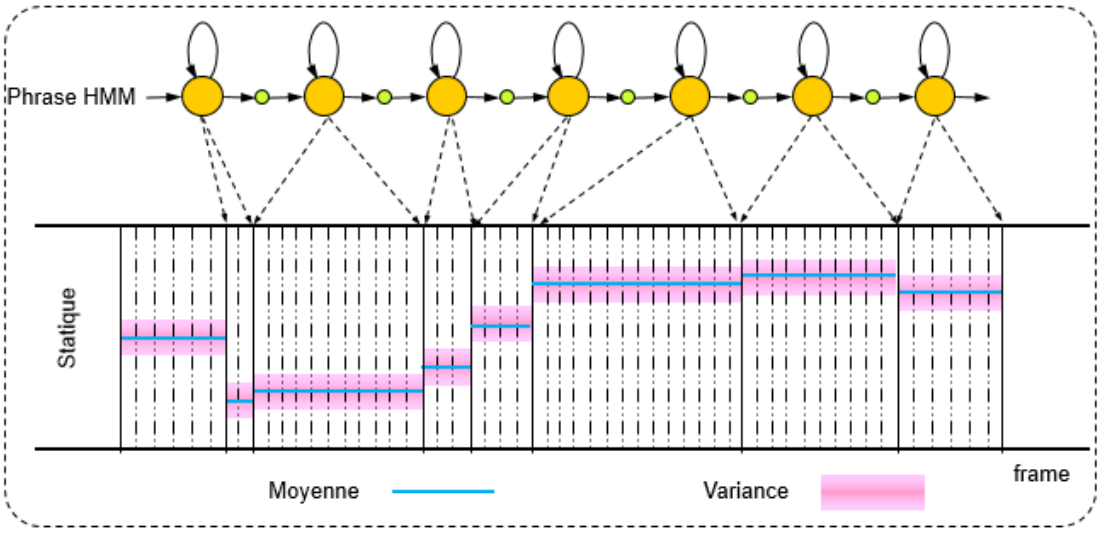


FIGURE 2.10 – Algorithme de génération des paramètres. [1]

L'algorithme de génération des paramètres maximise la probabilité d'observation (la moyenne de la gaussienne) pour chaque état en considérant uniquement les paramètres statiques. Cette méthode résulte en un vecteur de paramètres formé par une séquence des moyennes : une trajectoire des paramètres constante par morceaux qui change de valeur à chaque transition d'état ce qui rend la parole synthétisée discontinue. Une solution à ce problème est de considérer que pour chaque état, le vecteur des paramètres contient

1. http://hts.sp.nitech.ac.jp/archives/2.2beta/\acrshort{hts}_Slides.zip

non seulement les coefficients spectraux mais aussi leurs dérivées premières et dérivées secondes.

La méthode précédente ne tient pas compte des propriétés statistiques des paramètres statiques. Afin de pouvoir synthétiser une parole proche de la parole naturelle, il est nécessaire de suivre la vitesse de changement des paramètres statiques en utilisant les propriétés statistiques des coefficients delta et delta-delta (respectivement dérivées premières et secondes des paramètres acoustiques) (figure 2.11).

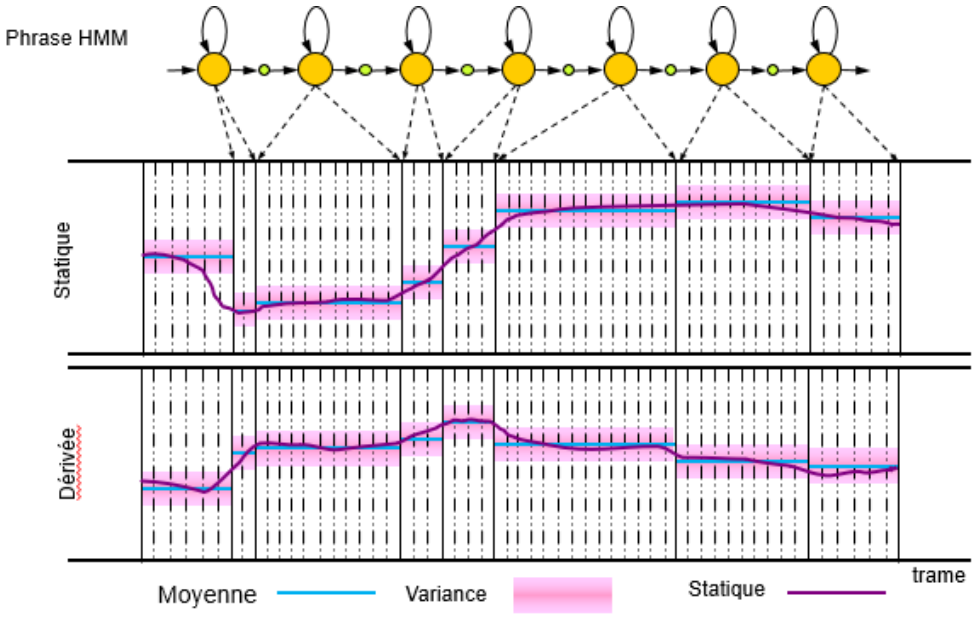


FIGURE 2.11 – Algorithme MLPG. [2]

La figure 2.12 présente les spectres calculés à partir des vecteurs des coefficients spectraux générés sans et avec les coefficients dynamiques. Cette représentation montre qu'une séquence de spectres variant sans discontinuité peut être obtenue en prenant en compte les caractéristiques dynamiques des coefficients statiques.

2. http://hts.sp.nitech.ac.jp/archives/2.2beta/\acrshort{hts}_Slides.zip

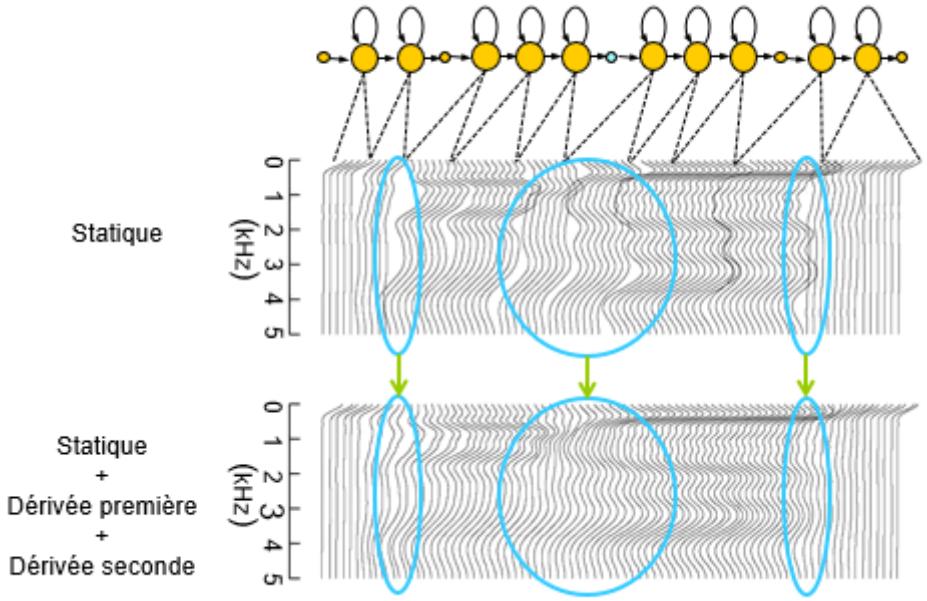


FIGURE 2.12 – Paramètres spectraux avec et sans utilisation des coefficients dynamiques [3]

2.3.6 Modélisation de F0

La représentation des valeurs observées de la fréquence fondamentale F0 est composée de valeurs à distributions continues pour les zones voisées et une valeur discrète pour les zones non voisées. Un HMM standard ne peut pas être appliqué dans ce cas, car F0 n'est pas définie dans les zones non voisées. Ainsi une nouvelle structure de HMM a été mise en place pour représenter statistiquement les observations de F0. Il s'agit de MSD (Multi-Space Probability Distribution)-HMM (Tokuda *et al.*, 2000).

MSD-HMM présente une structure à plusieurs espaces de distributions, dont chaque espace possède sa propre dimension. Certains peuvent être unidimensionnels, certains autres peuvent être bidimensionnels et d'autres espaces peuvent être de dimension zéro, ce qui représente une information discrète. La figure 2.13 montre la modélisation de F0 en utilisant MSD-HMM. Chaque état constituant le phrase HMM possède les probabilités d'observations voisées vs. non-voisées et une distribution continue pour les observations

3. http://hts.sp.nitech.ac.jp/archives/2.2beta/\acrshort{hts}_Slides.zip

voisées. Pareillement à la modélisation des paramètres du spectre, les coefficients de delta et delta delta ont été utilisés pour la modélisation de F0. En utilisant MSD-HMM, chaque ensemble de coefficients possède une distribution voisée et une autre non-voisée. L'utilisation des coefficients dynamiques permet de générer une trajectoire de F0 plus proche de celle obdervée pour la parole naturelle.

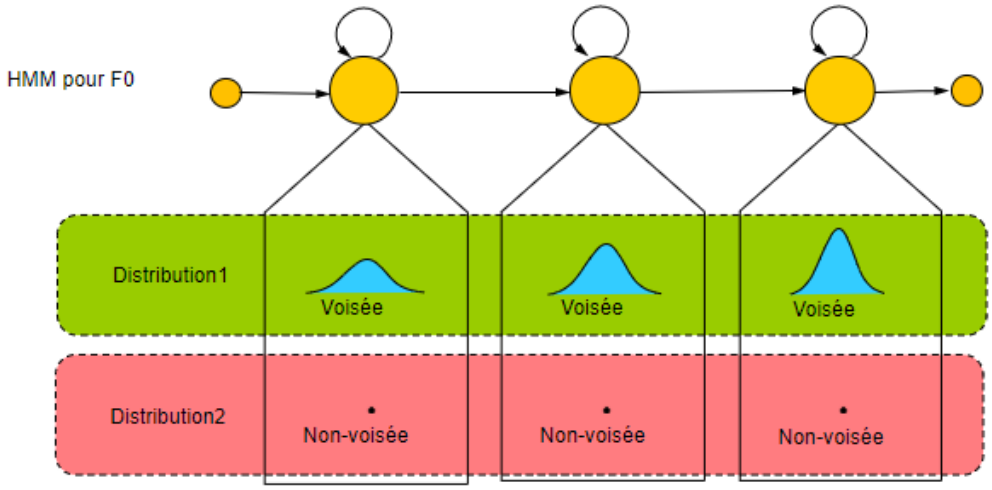


FIGURE 2.13 – Modélisation de F0 par MSD-HMM. [4]

2.3.7 Avantages de l'approche HMM

Comparée aux différentes approches de synthèse de parole, la synthèse paramétrique statistique présente un ensemble d'avantages dus principalement à l'utilisation des paramètres pour décrire les signaux et à l'utilisation des modèles statistiques.

Transformation des voix et émotions

L'approche paramétrique statistique de synthèse de parole permet de modifier les caractéristiques de la voix générée par transformation des paramètres des modèles. Cet aspect rend la modification des styles et des émotions possibles. Quatre techniques ont été étudiées dans la littérature ; par adaptation, interpolation, vecteur propre et régression multiple.

4. http://hts.sp.nitech.ac.jp/archives/2.2beta/\acrshort{hts}_Slides.zip

Adaptation : Cette technique a été tout d'abord développée pour la reconnaissance de la parole afin d'ajuster des modèles acoustiques à un locuteur ou à un environnement dans le but d'améliorer la précision de la reconnaissance (Leggetter et Woodland, 1995). La technique d'adaptation a été ensuite introduite dans la synthèse de parole par HMM par (Tamura *et al.*, 2001) pour pouvoir concevoir un système de synthèse dépendant du locuteur avec une taille réduite des données. L'adaptation peut être réalisée par deux méthodes : MAP (Maximum a posteriori) (Gauvain et Lee, 1994) ou MLLR (Maximum likelihood linear regression) (Leggetter et Woodland, 1995).

La méthode MAP est basée sur une connaissance de la distribution des paramètres des modèles. Ceci permet de mieux réaliser l'adaptation quand la quantité des données d'adaptation est petite. L'inconvénient principal de cette méthode est que l'adaptation se fait par distribution, ainsi si les données sont éparpillées, plusieurs modèles ne seront pas traités. Par conséquent, les caractéristiques de la parole synthétisée alterneront entre celle du locuteur original et celle de la parole cible.

L'adaptation peut aussi être réalisée en utilisant la méthode MLLR. Un ensemble de transformations linéaires est utilisé pour convertir les modèles en des nouveaux modèles adaptés. Les distributions des états des modèles sont regroupées en utilisant des arbres de décision de régression. En changeant la taille de l'arbre selon les données d'adaptation disponibles, la complexité et la généralisation de l'adaptation peut être contrôlée. (Yamagishi et Kobayashi, 2007) ont appliqué cette méthode à la synthèse de parole par HMM, la technique résultante est appelée AVSS (Average Voice Speech Synthesis). Elle permet de synthétiser de la parole avec les caractéristiques souhaitées même en utilisant une quantité réduite des données pour l'adaptation.

Interpolation : La technique d'interpolation permet de synthétiser de la parole sans passer par l'apprentissage des caractéristiques de la parole cible. En l'appliquant à la synthèse de parole par approche paramétrique statistique, les paramètres des HMM sont interpolés parmi les ensembles représentant des HMM (Tachibana *et al.*, 2005).

Par vecteur propre : Eigenvoce : La méthode d'adaptation permet en fait d'imiter les caractéristiques et les styles des données d'adaptation, ainsi l'adaptation ne peut pas être réalisée s'il n'y a pas des données correspondantes aux caractéristiques cibles. Cependant la technique d'interpolation rend l'obtention des nouvelles voix avec des nouvelles caractéristiques possibles ceci en modifiant le rapport entre les représentants des ensembles des HMM. Dans ce cas, si le nombre des ensembles des représentants

des HMM augmente afin d'améliorer la représentativité, la détermination du ratio des caractéristiques cibles devient difficile. Pour remédier à ce problème, la technique du vecteur propre (Kuhn *et al.*, 2000) a été appliquée à la synthèse de parole par HMM (Shichiri *et al.*, 2002).

Un vecteur spécifique à chaque locuteur est composé en concaténant la moyenne des vecteurs des distributions des états de sorties dans chaque ensemble des modèles. En appliquant l'Analyse en Composantes Principales (ACP) aux ensembles des vecteurs spécifiques, les vecteurs et valeurs propres sont obtenus. Seuls les vecteurs propres d'ordre inférieur sont gardés. Les premiers K vecteurs propres pondérés permettent d'obtenir les caractéristiques de la voix à générer ce qui permet de faire la construction d'un nouvel HMM. Cette méthode présente l'avantage de réduire le nombre des paramètres à contrôler mais rend le contrôle des caractéristiques de la voix synthétisée difficile.

Régression multiple : Pour résoudre ce problème, une approche de régression multiple (Fujinaga *et al.*, 2001) a été appliquée à la synthèse de parole par HMM (Masuko *et al.*, 2004) afin de contrôler les caractéristiques de la voix synthétisée. Dans cette méthode, le vecteur des distributions des états est contrôlé par un vecteur dont les composantes sont des caractéristiques possibles de la voix synthétisée (style, expression, émotion, genre). Ainsi, il suffit de préciser le vecteur de contrôle correspondant aux caractéristiques de la voix cible pour la synthétiser.

La combinaison de ces méthodes permet de générer des voix avec différents caractéristiques, styles et émotions sans avoir recours à des quantités importantes des données (Tachibana *et al.*, 2008).

L'espace acoustique

La synthèse par les méthodes de concaténation se base sur la sélection des unités à concaténer selon des critères bien définis (le coût de concaténation et coût de cible). La qualité du résultat obtenu est étroitement liée aux unités de parole présentes dans le corpus de parole, s'il n'y a pas suffisamment d'exemples pour certaines unités, ceci peut dégrader la qualité du signal généré. La synthèse par approche paramétrique statistique fait face à ce problème, car elle se base sur l'utilisation des statistiques pour générer les signaux de parole. Ainsi l'approche paramétrique statistique a recours à l'utilisation d'arbres de décisions pour partager les paramètres des modèles selon les contextes.

Empreinte réduite

Les approches de synthèse par concaténation exigent le stockage des unités de parole, ainsi la taille mémoire nécessaire est un facteur critique pour ces méthodes de synthèse. La synthèse paramétrique statistique ne présente pas ce problème, car seuls les statistiques des modèles acoustiques sont conservées au lieu des unités de parole (Zen *et al.*, 2007). Ceci rend possible l'utilisation de cette approche de synthèse dans les applications embarquées (Kim *et al.*, 2006).

Séparation de la modélisation de la durée de celle du spectre

La synthèse de parole par approche paramétrique statistique est basée sur la modélisation source/filtre de parole, ainsi la durée, l'excitation (fréquence fondamentale) et les paramètres spectraux sont contrôlés et modifiés séparément. Le modèle de durée consiste à estimer la durée de chaque état en nombre de trame. En utilisant les HMM standards, cette durée est estimée à une seule trame pour tous les états ce qui n'est pas consistent avec la parole naturelle. Aussi, un modèle explicite de durée a été introduit. Les différents modèles de durée, des paramètres d'excitation et du spectre interagissent à travers la structure du modèle. Toutefois, les facteurs contextuels affectant la durée sont différents de ceux affectant la fréquence fondamentale et les paramètres du spectre, c'est pour cela que les arbres de décision pour le partage des paramètres contextuels sont spécifiques à chaque modèle.

2.3.8 Adaptation aux autres langues

HTS a été adapté à plusieurs langues. Pour ce faire, il faut adapter l'ensemble des descripteurs contextuels aux particularités de la langue. Pour l'allemand, un descripteur à propos la position du phonème a été ajouté, il permet de déterminer si le phonème est le seul composant de la syllabe, sinon il fait partie de l'amorce, du noyau ou de la coda (Krstulovic *et al.*, 2007). Pareillement pour le basque (Erro *et al.*, 2010), un descripteur décrivant la position du phonème entre deux silences a été ajouté, et le descripteur indiquant la position du phonème dans la syllabe n'a pas été utilisé pour l'espagnol (Banos *et al.*, 2008).

Les descripteurs à l'échelle de syllabes ont été modifiés pour certaines langues. Principalement ceux liés à l'accentuation du mot (l'accent lexical et l'accent tonique). Certaines langues ne prennent pas en considération l'accent lexical, c'est le cas du suédois (Lundgren,

2005) et le portugais brésilien (Maia *et al.*, 2003) ne prend pas en compte l'utilisation de l'accent tonique. Le finnois n'utilise aucun descripteur pour l'accentuation (Silén *et al.*, 2010). Pour le français, (Le Maguer *et al.*, 2013) l'accentuation a été remplacée par la proéminence de la syllabe, ceci en utilisant les règles décrites dans (Simon *et al.*, 2008).

Pour le japonais, les modifications à l'échelle de la syllabe ont pris une autre dimension. La syllabe a été remplacée par une autre unité : la more (Labrune, 2006). La more est une unité phonologique permettant de quantifier la durée de la syllabe. La more peut être calculée sur le noyau et la coda ou bien sur le noyau seul selon les langues. Dans le cas du japonais, une voyelle courte implique que la syllabe est monomoraïque ; une syllabe contenant une diphtongue sera bimoraïque. La coda n'est pas prise en compte dans le calcul pour le Japonais (Le Maguer *et al.*, 2013).

A l'échelle du mot, le descripteur faisant référence à l'étiquette grammaticale du mot, n'a pas été utilisé pour plusieurs langues, il a été introduit principalement pour le français (Le Maguer *et al.*, 2013), l'allemand (Krstulovic *et al.*, 2007), le portugais brésilien (Maia *et al.*, 2003) et le basque (Erro *et al.*, 2010). Ce dernier a simplifié ce descripteur en lui attribuant deux valeurs seulement : le mot est fonctionnel ou lexical. L'échelle de la phrase a été remplacée par la notion de syntagme pour le français (Le Maguer *et al.*, 2013). Il s'agit d'un ensemble de mots formant une seule unité catégorielle et fonctionnelle. Ils constituent ensemble une unité sémantique, mais dont chaque constituant, parce que dissociable (contrairement au mot composé), conserve sa signification et sa syntaxe propre. De même, le descripteur ToBI (Silverman *et al.*, 1992), a été utilisé pour l'allemand (Krstulovic *et al.*, 2007) et l'anglais (Black *et al.*, 2007). Un descripteur faisant référence au type de la phrase a été ajouté pour le japonais pour indiquer s'il s'agit d'une question par exemple. Finalement, à l'échelle de l'énoncé, le basque (Erro *et al.*, 2010) a introduit un descripteur relatif à l'émotion.

Pour certaines langues, des modifications ont été apportées au vocodeur afin d'améliorer la qualité de la parole générée. C'est le cas de (Abdel-Hamid *et al.*, 2006) qui a apporté des modifications à HTS en l'appliquant à la langue arabe. Le but de ce travail était la modification des paramètres du signal de parole et l'utilisation d'un nouveau modèle d'excitation.

2.3.9 Discussion

Avec l'approche de synthèse de parole par HMM, les signaux générés n'ont pas encore atteint la même qualité et l'aspect naturel obtenu par la synthèse par sélection d'unités (Zen *et al.*, 2009). La dégradation de la qualité est due principalement à trois facteurs. Tout d'abord, l'utilisation du vocodeur. En effet, les vocodeurs qui sont basés sur une excitation mixte (Kawahara *et al.*, 1999) (périodique et apériodique) améliorent la qualité mais elle reste toujours moins bonne que celle obtenue avec l'approche par sélection des unités qui ne sont pas modifiées par un traitement ou un codage. Un autre inconvénient de cette approche est le lissage qui amène à générer une parole "étouffée". La méthode de variance globale GV (Global Variance) (Toda et Tokuda, 2007) a été utilisé pour remédier à ce problème.

Un troisième facteur affectant la qualité de la parole générée est le modèle acoustique qui assure le passage des descripteurs contextuels aux paramètres acoustiques. Dans l'approche standard de synthèse de parole par HMM, la relation entre les paramètres contextuels et les paramètres acoustiques est modélisée par les arbres de décision qui regroupent les contextes associés à des distributions similaires. Chaque groupement partage la même fonction de densité qui servira à la génération des paramètres acoustiques pendant la phase de synthèse. Les regroupements sont déterminés grâce à des arbres de décision. ((Zen, 2013), (Zen *et al.*, 2013), (Watts *et al.*, 2016)) ont montré que les arbres de décisions présentent des limitations qui peuvent affecter la qualité de la parole synthétisée. Il s'agit d'une architecture peu profonde voire superficielle qui ne peut pas modéliser des fonctions complexes ainsi que toutes les dépendances entre les descripteurs contextuels et les paramètres acoustiques. Considérant la taille de l'ensemble des descripteurs, la taille de l'arbre de décision devrait être très grande afin de pouvoir les modéliser. De même, pendant l'apprentissage, les données d'apprentissage sont divisées et partagées en groupes, chaque groupe possède ses propres paramètres. Ces facteurs peuvent affecter l'estimation des distributions des paramètres qui seront utilisées plus tard pour la prédiction des paramètres acoustiques correspondants au texte à synthétiser.

2.4 Synthèse de la parole par les réseaux de neurones

(Zen *et al.*, 2013) ont proposé l'utilisation des réseaux de neurones DNN (Deep Neurel Network) comme un choix alternatif pour remplacer les arbres de décisions. En effet,

depuis quelques années, les réseaux de neurones ont été utilisés pour la modélisation des signaux de parole, principalement pour la reconnaissance automatique de la parole. L'utilisation de ces architectures décrites comme profondes, présente plusieurs avantages tels que l'apprentissage des données sans les diviser en groupes et la capacité de modéliser des fonctions complexes. Cette section décrit l'utilisation des réseaux de neurones en synthèse paramétrique de la parole.

2.4.1 Aspect général des réseaux de neurones

Le cerveau humain organise hiérarchiquement les idées et les concepts, un réseau de neurones est construit. D'après la neurophysiologie, il contient près de 100 milliards cellules (neurones). Elles sont toutes interconnectées via les dendrites (en entrée) et les synapses (en sortie), formant ainsi un réseau de l'ordre de des millions de milliards de connexions (Gurney, 2014). Chaque neurone reçoit des milliers de connexions de la part des autres neurones, ainsi chaque cellule reçoit constamment des nombreux signaux. Les signaux sont ou bien sommés ou intégrés tous ensemble. Si le signal résultant dépasse un certain seuil, la cellule va l'ignorer ou elle envoie une impulsion de tension. Le signal sera ensuite transmis aux autres cellules par l'intermédiaire de l'axone. Une partie des signaux d'entrées produit un effet inhibiteur et empêche la génération des impulsions ; contrairement aux signaux d'excitation qui favorisent la génération des impulsions. La nature du traitement effectué par chaque neurone dépend de la nature du signal d'entrée ainsi que de la puissance des connexions avec les autres neurones.

Une cellule neuronale artificielle est appelée perceptron ; la synapse est modélisée par une valeur numérique appelé poids. Chaque entrée est multipliée par le poids correspondant avant d'être envoyée vers le cœur de la cellule (figure 2.14).

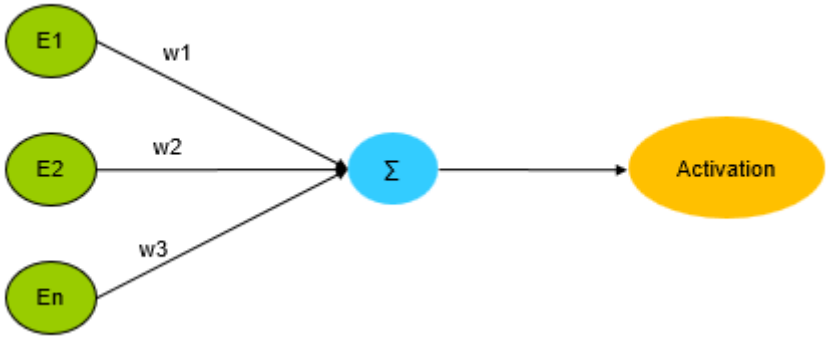


FIGURE 2.14 – Un perceptron

La sortie du perceptron est binaire, elle est obtenue en comparant la somme des entrées multipliées chacune par les poids à un certain seuil (Nielsen, 2015). Le seuil est une valeur numérique qui représente une caractéristique de la cellule. Un réseau de neurone artificiel est formé d'un ensemble des couches de perceptrons avec généralement une structure d'au moins trois niveaux : chaque signal passe par l'entrée puis traverse les couches cachées pour arriver en sortie. Dans le réseau de la figure 2.15, chaque unité (cellule) est connectée à toutes les cellules de la couche inférieure, et à toutes les cellules de la couche supérieure. Les connexions entre les cellules sont caractérisées par leur poids.

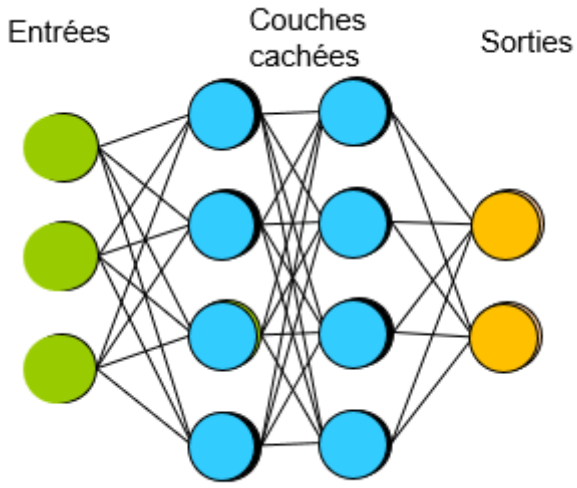


FIGURE 2.15 – Réseau de neurones artificiels

2.4.2 Synthèse de la parole par réseaux de neurones

(Zen *et al.*, 2013) a introduit l'utilisation des réseaux de neurones dans l'approche de synthèse paramétrique. Ceci en remplaçant les arbres de décisions par un réseau de neurones. Ainsi, le passage des descripteurs contextuels aux paramètres acoustiques est assuré par les réseaux de neurones (figure 2.16).

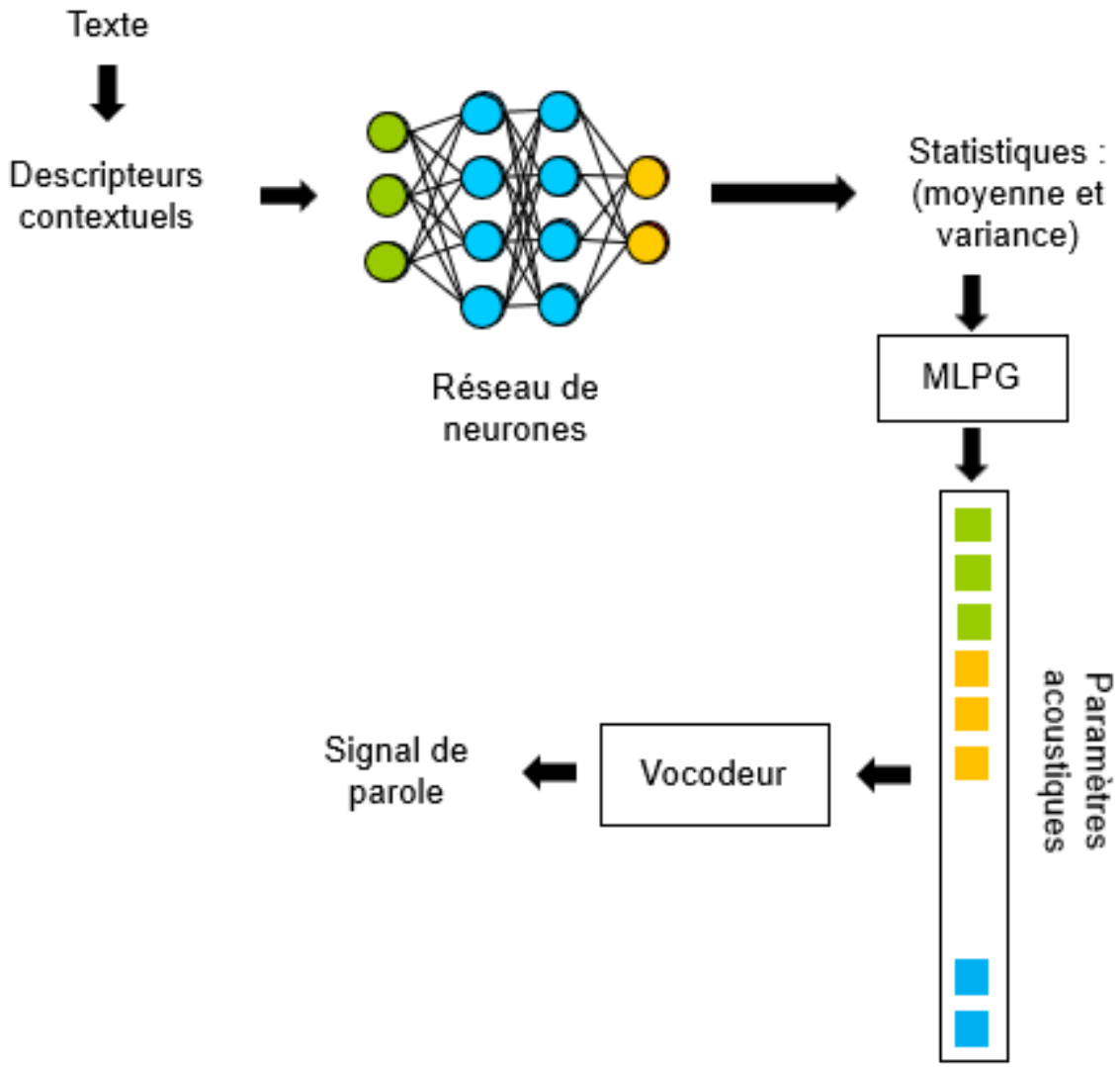


FIGURE 2.16 – Aspect général de la synthèse par DNN

2.4.3 Aspect général

Comme pour l'approche paramétrique statistique, l'entrée des systèmes de synthèse est un texte écrit. Le texte subit ensuite l'ensemble des traitements nécessaires afin d'extraire les descripteurs contextuels correspondant à la séquence des phonèmes. Typiquement, il s'agit des mêmes informations que celles extraites pour la synthèse par HMM (Tokuda *et al.*, 2002). Ces informations sont codées sous forme numérique pour être fournies en entrée du réseau de neurones. À ces informations s'ajoutent les informations des durées. À la sortie des réseaux de neurones, les statistiques des paramètres acoustiques sont obtenues. Un algorithme de génération des paramètres est utilisé pour générer le vecteur des paramètres acoustiques (les coefficients spectraux, les paramètres d'excitation et de voisement et non voisement). Ces paramètres sont ensuite insérés dans un vocodeur pour générer le signal de parole correspondant au texte de l'entrée.

Différentes architectures ont été utilisées pour implémenter la synthèse de parole par les réseaux de neurones. (Zen *et al.*, 2013) utilise un réseau de neurones (perceptrons multicouches) standard. Le vecteur d'entrée contient les réponses binaires aux questions sur le contexte linguistique et des valeurs numériques (telle que la position du phonème dans la syllabe,...), la position relative de la trame courante dans le phonème et les durées des phonèmes et des états. Les vecteurs des entrées sont convertis en des vecteurs de sorties par l'intermédiaire d'un réseau de neurones à propagation en avant. Le vecteur des sorties contient les paramètres spectraux (et leurs dérivées) et les paramètres d'excitation (et leurs dérivées).

Les paramètres des DNN sont appris en utilisant des paires des paramètres d'entrée et de sortie extraits à partir des données d'apprentissage. Pour les expériences, une base de données des phrases en anglais a été utilisée pour comparer les performances des HMM à celles des DNN. Les HMM utilisaient des arbres de décision et 2554 questions pour leur construction. Le vecteur d'entrée des DNN comportait 381 paramètres (dont 342 paramètres binaires qui sont les réponses aux questions contextuelles, les 39 restants sont des paramètres numériques de type positions et durée du segment utilisé).

Pour la prédiction de F0, (Zen *et al.*, 2013) s'est basé sur les valeurs continues de F0 accompagnées d'un modèle explicite de voisement, la valeur de log $F0$ a été interpolée dans les zones non voisées. Pendant l'apprentissage, les silences ont été éliminés afin de réduire le coût du calcul. L'erreur quadratique entre les paramètres d'entrée et les paramètres prédits a été réduite en utilisant l'algorithme SGD (Stochastic Gradient Descent). Les

résultats des évaluations objectives ont montré que les DNN ont surpassée les HMM au niveau de la prédiction des paramètres de l'apériodicité et la classification des zones voisées et non voisées. Les paramètres spectraux sont mieux prédits avec des réseaux profonds (plusieurs couches cachées). Cependant une meilleure prédiction de log F0 a été obtenue avec les HMM, ceci peut être dû à toutes ces trames non voisées qui ont été interpolées et modélisées comme étant des trames voisées (Zen et Senior, 2014).

2.4.4 Utilisation des architectures DNN dans la synthèse de parole

(Fan *et al.*, 2014) a montré que l'utilisation des DNN standards pour la synthèse de parole, n'a pas encore atteint le niveau souhaité en ce qui concerne l'aspect naturel de la parole générée. Cette dernière présentait des discontinuités. Ceci est dû au fait que chaque trame est modélisée séparément de ses trames voisines et que les réseaux de neurones utilisés nécessitent l'association des paramètres dynamiques aux paramètres statiques afin de garantir le lissage de la parole synthétisée. Dans cet article, les auteurs proposent la synthèse de la parole à partir d'un texte par RNN (Recurrent Neural Network), par LSTM (Long Short Term Memory) et par BLSTM (Bidirectional Long Short Term Memory).

Profiter des avantages de RNN et LSTM, et leur nature séquentielle en particulier, peut mettre fin à ces problèmes. L'algorithme BPTT (Back Propagation Through Time) est utilisé pour faire l'apprentissage des RNN et de LSTM bidirectionnelles et le calcul des gradients des poids des RNN est calculé par rapport à l'énoncé tout entier (avantage des RNN par rapport aux DNN). Pendant la phase de synthèse et après la conversion du texte en une suite de vecteurs des paramètres des entrées, les vecteurs des sorties sont obtenus en utilisant un LSTM bidirectionnelles-RNN. Pour synthétiser de la parole par HMM ou par DNN, les paramètres dynamiques sont indispensables pour générer les trajectoires des paramètres de la parole. Cependant, les RNN ne font pas recours à ces informations dynamiques, et n'ont comme sortie que les paramètres statiques, ceci est dû à la considération de la nature séquentielle du problème par les RNN : le module de génération des paramètres est intégré dans LSTM bidirectionnelles-RNN.

Afin d'avoir le même nombre des paramètres pour les trois systèmes, une ou deux couches supérieures de DNN ont été remplacées par BLSTM-RNN. Les résultats de l'évaluation objective ont montré que le nombre de couches cachées n'a pas beaucoup influencé sur

les mesures objectives : F0 RMSE, erreur de permutation des voisement/non voisement et la distance normalisée du spectre. Comparé aux HMM et DNN, les systèmes à base de LSTM bidirectionnelles-RNN ont une meilleure performance et sont en égalité avec les DNN au niveau de RMSE. (Zen et Sak, 2015) ont montré que l'utilisation des LSTM bidirectionnelles, peut augmenter le temps de réponse du système. Dans cet article, l'accent est mis sur l'utilisation des LSTM dans la synthèse de la parole par approche paramétrique statistique et son effet sur le temps de réponse. L'auteur a proposé une architecture de synthèse de parole en streaming avec un temps de réponse réduit, ceci en utilisant LSTM-RNN unidirectionnel. En utilisant une architecture séquentielle à base de LSTM bidirectionnelle, une transition lisse entre les trames voisines est garantie mais le temps de réponse augmente largement, ceci est une conséquence de la propagation dans les deux sens (avant et arrière) ; afin de prédire la séquence de la première trame, les entrées de la dernière trame doivent se propager à travers le réseau. Ce problème disparaît avec les LSTM-RNN unidirectionnel, car la propagation en avant est faite d'une manière synchrone et en streaming. Pour évaluer les performances de cette architecture, une comparaison est faite entre les architectures suivantes :

— DNN (standard : des multi-couches de perceptrons)

— LSTM-RNN unidirectionnelle

Cette comparaison avait pour but l'évaluation de l'effet de l'utilisation de l'algorithme (de lissage) de génération des paramètres. Les DNN et LSTM-RNN sont appris avec et sans caractéristiques dynamiques. L'évaluation subjective à base des tests de préférence a montré que le lissage des trames est essentiel dans le cas d'utilisation des DNN. LSTM-RNN unidirectionnelle produit une meilleure qualité de parole comparée à celle par DNN et le lissage est moins pertinent dans ce cas de LSTM-RNN mais il a été jugé utile. Le lissage par caractéristiques dynamiques et celui par la couche récurrente donnent des résultats similaires. Les résultats des tests MOS montrent une meilleure efficacité de la modélisation acoustique par LSTM-RNN unidirectionnelle par rapport à celle par DNN.

2.5 MERLIN

MERLIN est un système de synthèse de parole reposant sur l'approche paramétrique et utilisant les réseaux de neurones (Wu *et al.*, 2016). Le système est basé sur la prédiction des paramètres acoustiques à partir des paramètres contextuels en utilisant les réseaux

de neurones. Cet outil propose l'utilisation de différentes architectures (LSTM BLSTM RNN GRU) pour modéliser les paramètres extraits à partir des signaux de parole naturelle pendant l'apprentissage et pour le passage des informations contextuelles aux paramètres acoustiques pendant la phase de synthèse. La phase de traitement du texte d'entrée est assurée par un outil externe (par exemple Festival (Black *et al.*, 2002) ou Osian). Pour la génération des signaux de parole, MERLIN est compatible avec les vocodeurs STRAIGHT (Kawahara *et al.*, 1999) et WORLD (Morise *et al.*, 2016). Pour assurer la bonne mise en forme des entrées des réseaux de neurones, MERLIN propose deux types de normalisation des données :

— Min-Max : normalisation des données dans la plage [0.01 0.99], principalement pour les paramètres contextuels.

— Variance moyenne : normalisation des données pour obtenir une moyenne nulle et la variance unitaire. Ce type de normalisation est destiné aux paramètres acoustiques.

MERLIN a été utilisé pour l'ensemble des expériences de synthèse de parole par DNN présentées dans la suite de ce document.

2.6 Conclusion

Ce chapitre s'est intéressé à une approche particulière de synthèse de parole à partir d'un texte : l'approche paramétrique. Elle est basée sur une description du signal par un ensemble de paramètres acoustiques qui rend possible la reconstruction du signal en utilisant ces paramètres. De même, la synthèse paramétrique qualifie chaque segment du texte par un ensemble de descripteurs contextuels qui comporte tous les facteurs pouvant affecter la prononciation du son correspondant. La synthèse par approche paramétrique statistique basée sur les HMM a été décrite. Le processus de synthèse par le système HTS a été ensuite présenté. La dernière partie du chapitre s'intéressait à la synthèse de parole par les réseaux de neurones ; leur principe et comment ils sont utilisés en synthèse de parole à partir d'un texte.

Chapitre 3

Adaptation de la synthèse paramétrique à la langue arabe

Sommaire

3.1 Introduction

Ce chapitre s'intéresse à la synthèse paramétrique de la parole arabe à partir d'un texte. Une première partie est dédiée à la présentation des particularités de la langue arabe, en écriture et en phonologie. La suite présente l'état de l'art de la synthèse de la parole arabe en utilisant les différentes approches citées dans les deux premiers chapitres.

Les deux sections suivantes présentent deux aspects de mon travail de thèse pour l'application de la synthèse paramétrique à la langue arabe : d'une part nos propositions de choix des unités de modélisation, et d'autre part notre proposition d'adaptation des descripteurs contextuels aux particularités de la langue arabe.

3.2 La langue arabe

La langue arabe fait partie des langues sémitiques (ougaritique, phénicien, araméen, hébreu, et arabe) et est parlée par plus de 530 millions de locuteurs dans le monde. Trois catégories sont distinguées : l'arabe classique, l'arabe standard MSA (Modern Standard Arabic) et l'arabe dialectal. L'arabe classique est défini comme la langue formelle parlée pendant l'époque de premières rédactions du Coran. L'arabe dialectal est lié à l'origine de la personne et varie selon les pays arabophones ou même selon les régions dans un pays. Dans ce présent travail, seul l'arabe standard a été considéré. Il représente une forme de langue commune à tous les locuteurs, il est enseigné à l'école. L'arabe standard s'écrit de droite à gauche, alors que les nombres sont écrits de gauche à droite.

3.2.1 Écriture

La langue arabe est caractérisée par un alphabet qui contient un ensemble de 28 lettres dont 25 sont des consonnes, les 3 restantes sont des voyelles longues. Des signes diacritiques indiquent l'identité des voyelles courtes :

La voyelle /a/ : Elle est représentée par le signe "fatha" au-dessus de la consonne : دَ /da/.

La voyelle /u/ : Elle est représentée par le signe "damma" au-dessus de la consonne : دُ /du/.

La voyelle /i/ : Elle est représentée par le signe "kasra" au-dessous de la consonne : دِ /di/.

Il existe aussi d'autres signes diacritiques :

Le tanwin : Le signe de tanwin est ajouté à la fin des mots indéterminés, il correspond à la prononciation du son /n/ à la fin du mot. (Kouloughli, 2007) : ceci consiste à doubler un des signes diacritiques déjà mentionnés :

- بٌ : /bun/
- بًا : /ban/
- بٍ : /bin/

Le "sokun" : Ce signe est utilisé pour indiquer l'omission d'une voyelle. Il s'agit d'un petit cercle au-dessus de la consonne : لَكِنْ /laki**n**/ "mais"

La "shadda" : Ce signe en forme qui a la lettre "w' : /ـّـ/ est utilisé pour distinguer les consonnes géminées des consonnes simples : نَزَّلَ /nazzala/ "faire descendre" (voir section 3.2.2).

Généralement, dans les textes modernes, les signes diacritiques correspondant aux voyelles courtes ne sont pas représentées.

3.2.2 Phonologie

Différentes études approfondies ont décrit un ensemble de caractéristiques phonologiques du MSA (Al-Ani, 1970) :

Les voyelles longues

La langue arabe présente deux types de voyelles : courtes et longues (Newman, 2002). En écriture, contrairement aux voyelles courtes, les voyelles longues sont toujours indiquées par les graphèmes suivants :

- /a :/ par [ا]

- /u :/ par [و]
- /i :/ par [ي]

Acoustiquement, une voyelle longue est environ deux fois plus longue que la voyelle courte correspondante (Khouja et Zrigui, 2005). En plus, si une voyelle courte est remplacée par une voyelle longue, le sens du mot change. *Exemple* : هتف /hatafa/ qui veut dire "il a crié", avec une première voyelle courte /a/. Si celle-ci est remplacée par la voyelle longue correspondante /a :/, le mot devient هَاتف /ha :tafa/ qui veut dire "il a téléphoné".

Consonnes géminées

Les consonnes de la langue arabe peuvent exister sous deux formes : simple et géminée. En écriture, la consonne géminée est distinguée de son homologue simple par l'ajout d'un signe diacritique appelé "shadda" /ـّ/ au-dessus de la consonne concernée (Newman, 2002). Acoustiquement, une consonne géminée possède une durée supérieure (double) à celle de son homologue simple (Khouja et Zrigui, 2005). Le remplacement d'une consonne simple par la consonne géminée correspondante change le sens du mot. *Exemple :* دَرَسَ "darasa" qui veut dire "il a étudié", devient دَرَّسَ "darrasa" qui veut dire "il a enseigné".

Aspect emphatiques

Il s'agit de la pharyngalisation de certaines consonnes dans certains contextes (Halabi, 2015). Certaines voyelles peuvent présenter l'aspect emphatique si elles sont précédées ou suivies par une consonne emphatique (Watson, 2002). Les consonnes en arabe peuvent être regroupées en trois classes :

- Consonnes de nature emphatique : /$s^ʕ$/ ص ; /$t^ʕ$/ ط ; /q/ ق ; /$d^ʕ$/ ض ; /$ð^ʕ$/ ظ/ ; /ɣ/ غ ; /x/ خ.
- Consonnes qui peuvent être emphatiques dans certains contextes : /r/ ر ; /l/ ل.
- Les autres consonnes qui ne peuvent pas être emphatiques.

Classification des consonnes

Les consonnes peuvent être regroupées en différentes classes selon le mode et le lieu d'articulation :

Nasales : Elles sont produites en abaissant le voile du palais. Selon l'emplacement, deux possibilités de nasales : dentale-alvéolaire /n/ ن ou labiale /m/ م.

Plosives : une plosive est produite par la fermeture complète et momentanée du chenal expiratoire résultant du contact entre deux articulateurs.

- Les plosives voisées sont non-aspirées, tandis que les plosives non-voisées sont aspirées, à l'exception de /q/ ق.
- Au voisinage d'une voyelle frontale haute, telle que /i/ ou /i :/ (courte ou longue), la plosive /k/ ك est palatalisée.
- La bilabiale voisée /b/ ب est souvent dévoisée au voisinage d'un son voisé.

Vibrante : C'est un son voisé dentoalvéolaire /r/ ر.

Fricatives : Elles sont produites par resserrement du conduit vocal à un endroit variable. La friction peut être produite par différents organes et combinaisons (lèvres, langue, dents contre lèvres, dents contre langue, voile du palais...) : /tˤ/ ط ; /ðˤ/ ظ/ ; /z/ ز ; /ð/ ذ ; /ʕ/ ع ;/ɣ/ غ ; /H/ح ; /x/ خ ; /f/ ف ; /h/ ه ; /ʃ/ ش.

Affriquées : Elle est composée d'une phase occlusive (où le flux d'air est bloqué) suivie par une étape fricative (où l'air retenu est relâché pour passer par une ouverture plutôt étroite) : /ʤ/ ج.

Approximantes : Elles sont produite par un rapprochement modéré des organes phonateurs qui ne va pas jusqu'à produire le bruit caractéristique de friction des fricatives : /w/ و ; /y/ ي.

Latérale : Elle est formée par l'affaissement de l'avant de la langue et le contact de son dos avec le palais : /l/ ل.

Structure syllabique

MSA présente six structures syllabiques possibles ((Baloul, 2003), (Demri, 2016)) qui peuvent être classées en trois groupes : syllabe légère, lourde ou super-lourde :

Syllabe légère : Elle est de la forme de **"CV"**. Elle consiste en une consonne simple suivie par une voyelle courte. *Exemple :* CV : دَ /da/

Syllabe lourde : Deux formes sont possibles :

— une consonne simple suivie par une voyelle longue (**"CVV"** [1]). *Exemple :* بَا /ba :/.

— une voyelle courte entourée par deux consonnes simples **"CVC"**. *Exemple :* بَب /bab/.

Syllabe super-lourde : Elle peut exister sous trois formes :

— une consonne simple suivie par une voyelle longue et une consonne simple **"CVVC"**. *Exemple :* /nun/ : نُن.

— une consonne simple suivie par une voyelle courte et deux consonnes **"CVCC"**. *Exemple :* /barq/ : بَرق.

— une consonne simple suivie de deux voyelles et deux consonnes **"CVVCC"**. *Exemple :* /cha :bb/ : شَابّ

L'accent lexical

L'accent lexical ou encore accent du mot (Halpern *et al.*, 2009) correspond à la mise en relief d'une syllabe dans le mot. L'accent peut se manifester par une augmentation de l'intensité, de la longueur de la voyelle ou du pitch. En MSA, différentes règles ont été mises en place fin de prédire la position de l'accent dans le mot. (Koulouchli, 1976), s'est basé sur le fait que l'accent ne peut pas être au-delà des trois dernières syllabes pour décrire ses règles de prédiction de la position de l'accent lexical dans le mot :

- L'accent est porté par la dernière syllabe du mot si elle est super-lourde.
- Sinon, si l'avant dernière syllabe est lourde donc, elle porte l'accent du mot.
- Sinon l'avant avant dernière syllabe porte l'accent lexical du mot.

Ces règles diffèrent du principe des règles de (Al-Ani, 1970) qui considère que chaque syllabe présente trois niveaux d'accent (primaire, secondaire et troisième niveau) dont la présence et la position dépendent de la structure syllabique du mot :

- Si le mot ne contient que des syllabes légères, la première syllabe porte l'accent lexical du mot et les autres syllabes sont inaccentuées (accent de troisième niveau)

1. Pour toutes ces notations syllabiques, V indique une voyelle (courte/longues) et C une consonne (simple/géminée)

- Sinon, si le mot contient une syllabe lourde, celle-ci porte l'accent lexical, les autres syllabes sont inaccentuées et les syllabes lourdes à la fin du mot ne sont pas considérées.
- Sinon, si le mot contient deux syllabes lourdes ou plus, celle la plus proche de la fin du mot porte l'accent lexical du mot (l'accent primaire), celle la plus proche du début du mot porte l'accent secondaire et les autres sont inaccentuées.

La dernière étude le l'accent lexical en MSA a été réalisé par (Halpern *et al.*, 2009). L'étude était basée sur l'effet des dialectes sur la position de l'accent lexical et sa réalisation. Les règles sont présentées sont comme suit :

- Si la dernière syllabe est une syllabe lourde c'est elle qui porte l'accent lexical du mot.
- Sinon, si le mot est monosyllabique donc, cette unique syllabe porte l'accent lexical du mot.
- Sinon, si le mot contient deux syllabes, l'accent est porté par la première.
- Sinon, si le mot contient plus que deux syllabes et l'avant dernière syllabe est lourde, celle-ci porte l'accent lexical.
- Sinon l'accent est porté par l'avant avant dernière syllabe.

3.3 Synthèse de la parole arabe

L'intérêt à la synthèse de la parole arabe a suivi le développement des différentes techniques et approches de la conversion d'un texte écrit en une voix parlée. Les différentes méthodes de synthèse de la parole citées précédemment ont été appliquées et adaptées à la langue arabe.

3.3.1 Synthèse par règles

(Rajouani *et al.*, 1987) a appliqué la méthode de synthèse par règles à la langue arabe. Ses travaux étaient basés sur une segmentation à l'échelle des phonèmes et une investigation des particularités phonologiques et phonétiques de la langue arabe. Trois points ont été abordés : la conversion du texte en une séquence de phonèmes, les règles du traitement linguistique et la paramétrisation des sons arabes en fonction du synthétiseur utilisé.

La transcription des lettres en une séquence des phonèmes inclut l'ajout des marqueurs de l'accent lexical dont la position est prédite par un système de syllabification automatique et suivant les règles de (Al-Ani, 1970). En effet, le passage d'une transcription orthographique en une transcription phonétique est assuré par des tests arborescents tenant compte du contexte droit et gauche d'une fenêtre qui glisse tout au long du texte. Les règles du passage des phonèmes au signal de parole sont déterminées à plusieurs niveaux. Le niveau phonologique s'intéresse au traitement des voyelles, plus particulièrement les voyelles emphatiques (ce sont les voyelles en contact avec les consonnes emphatiques). Elles ont été gérées par un module spécifique basé sur des règles de propagation de la pharyghalisation tout au long de la parole continue.

De même, les consonnes géminées sont générées en utilisant les paramètres spectraux de la consonne simple correspondante et en augmentant la durée de l'état le plus stable de la consonne. Au niveau prosodique, la position de l'accent lexical est prédite. La même référence a montré que pour la langue arabe, si la voyelle est accentuée, la fréquence fondamentale augmente, l'amplitude s'accroît et le segment s'allonge. Dans la version décrite du système, le rythme est traité essentiellement pour les voyelles et tient compte de leur durée (courte ou longue), et de la nature de leur contexte, i.e., de présence de l'accent lexical et des consonnes adjacentes. Les indicateurs de la prosodie ont été placés manuellement en tenant compte de la structure syntaxique de la phrase à synthétiser.

Concernant la phase de synthèse, l'intérêt a été porté sur l'étude de l'évolution des trois premiers formants dans le temps, pendant la transition d'une consonne à une voyelle. Un ensemble de règles a été utilisé pour calculer les paramètres de contrôle qui sont envoyés au synthétiseur chaque 10 ms. Les formants cibles entre deux segments consécutifs sont interpolés en tenant compte des durées. L'amplitude de la source d'excitation et les formants sont calculés par les règles de transitions qui sont fonction de la classe des phonèmes des deux segments consécutifs et de la position du segment dans la séquence.

3.3.2 Synthèse par concaténation

Les méthodes de synthèse par concaténation ont été appliquées à la langue arabe en utilisant des segments de différentes tailles. (Chenfour *et al.*, 2000) a proposé l'utilisation de la di-syllabe comme unité de parole pour générer des signaux de parole à partir d'un texte écrit en utilisant la méthode de synthèse basée sur l'algorithme TD-PSOLA. Le choix de cette unité a été justifié par le fait qu'elle permet d'améliorer la qualité de la parole

générée par rapport à celle générée par une concaténation des diphones. (Chenfour *et al.*, 1997) définit la di-syllabe comme étant "une transition du noyau vocalique d'une syllabe vers le noyau vocalique de la syllabe suivante."
Les auteurs justifient ce choix en s'appuyant sur le fait que (Moulines *et al.*, 1990) a montré que les diphones ne sont pas suffisants pour "transporter" le mécanisme de coarticulation. Ainsi plus l'unité est longue, plus elle permet de synthétiser une meilleure qualité de parole. Comme le nombre de di-syllabes obtenues à partir des syllabes était trop élevé, des hypothèses ont été adoptées pour le réduire ; telle que la considération uniquement des di-syllabes contenant des voyelles courtes et la génération des voyelles longues en dupliquant des périodes stables des voyelles courtes ainsi que l'élimination des cas impossibles à avoir.

Pendant la phase de synthèse, lors de l'application de l'algorithme TD-PSOLA pour faire la concaténation des unités, une interpolation est utilisée afin d'éliminer les distorsions et les discontinuités spectrales. La discontinuité peut être due au fait que les deux di-syllabes à concaténer sont extraites des différents contextes.

(Ahmed, 2004) a proposé la concaténation d'unités de différentes tailles (des polyphones) selon des règles établies de syllabification et de transcription. Le système de synthèse de parole proposé repose sur une partie linguistique qui assure la transcription du texte à convertir en un signal de parole. Une simple concaténation des phonèmes ne mène pas à une construction correcte du signal de parole car les transitions entre les phonèmes responsables de la coarticulation ne sont pas prises en compte. Trois segments de différentes tailles ont été considérés : les triphones, les diphones et les phonèmes. L'utilisation d'unités de différentes tailles a permis d'avoir une meilleure qualité de parole. Dans ce même cadre, des règles de syllabification ont été établies :

- Lorsqu'une consonne est suivie d'une voyelle longue, les trois phonèmes constituent une unité acoustique.
- Lorsqu'une consonne est suivie d'une voyelle puis d'une consonne les deux premiers phonèmes constituent une unité acoustique.

Les travaux de (Halabi, 2015) se sont focalisés sur l'évaluation de l'ajout des indicateurs de l'accent lexical ou non lors de la production de la parole à partir du texte en utilisant le système de synthèse Innoetics basé sur une approche de concaténation (Chalamandaris *et al.*, 2013). Les règles de prédiction de la position de l'accent lexical citées dans (Halpern *et al.*, 2009) ont été utilisées. Trois systèmes ont été considérés :

- Système 1 : présence d'information sur l'accent lexical.

- Système 2 : absence d'information sur l'accent lexical.
- Système 3 : ajout des phrases sans sens.

Les différents systèmes ont été évalués en se basant sur des tests d'écoute de type MOS pour juger la qualité globale des signaux produits et de type DMOS pour évaluer le degré de dégradation des signaux synthétisés par rapport aux signaux naturels. Les résultats d'évaluation ont montré que tenir compte de l'accent lexical améliore la qualité de la parole synthétisée et réduit son degré de dégradation par rapport à la parole naturelle.

3.3.3 Synthèse par sélection d'unités

(Abdelmalek et Mnasri, 2016) ont mis en place un système de synthèse de parole arabe basé sur l'approche par sélection d'unités. Le processus de synthèse décrit passe par plusieurs étapes. Tout d'abord par le choix des phrases du corpus utilisé. Le choix tient compte des différents contextes possibles ainsi que des caractéristiques phonologiques de la langue arabe. Une fois le corpus prêt, il fallait choisir l'unité de parole qui sera utilisée pour la synthèse de parole. Après le calcul des coûts de concaténation, une sélection finale des candidats les plus proches avait lieu afin de générer le signal de parole. Pour les expériences, un ensemble de 30 phrases a été utilisé et les unités choisies sont les phonèmes et les syllabes : 546 phonèmes et 236 syllabes. La segmentation des phrases en phonèmes et en syllabes a été faite manuellement. Les signaux de parole obtenus ont été évalués subjectivement et objectivement et ont montré que la qualité et l'aspect naturel de la parole synthétisée présentent un score mos aux alentours de 4.1 sur 5.

3.3.4 Synthèse par approche paramétrique

Avec l'apparition de la méthode de synthèse par approche paramétrique statistique, (Abdel-Hamid *et al.*, 2006) a appliqué le système HTS à la langue arabe et il a introduit des modifications afin d'améliorer la qualité des signaux de parole synthétisés. Afin de supprimer le bourdonnement des signaux de parole synthétisés, la technique d'excitation multi-bandes MBE est utilisée, où la bande de fréquence de la parole est divisée en un certain nombre de sous-bandes. Un taux de voisement est estimé dans chaque sous-bande pour indiquer le degré de voisement. Les sources de voisement et de non-voisement sont mixés selon le taux de voisement. Et quand on augmente le nombre de paramètres mel-cepstraux, ils tendent à représenter les caractéristiques des formants.

Ainsi, dans la version modifiée du système, ces paramètres spectraux ont été remplacés par un plus grand nombre de paramètres d'enveloppe spectrale. Pendant la phase de synthèse, ces paramètres sont utilisés pour calculer les amplitudes des sinusoïdes des zones voisées, et permettent de filtrer la source d'excitation du bruit pour les zones non-voisées. Ainsi, les paramètres utilisés dans cette version modifiée du système de synthèse sont : les paramètres d'enveloppe spectrale, les bandes de voisement et la fréquence fondamentale auxquels s'ajoutent leurs dérivées premières et secondes pour l'apprentissage des modèles HMM dépendant du contexte. Les résultats obtenus ont montré que les modifications apportées au système ont améliorée la qualité de la parole synthétisée.

(Khalil et Adnan, 2013) ont aussi appliqué le système de synthèse de parole par HMM à la langue arabe. STRAIGHT (Kawahara *et al.*, 2001) a été utilisé pour garantir une meilleure représentation des paramètres acoustiques. Un corpus segmenté à l'échelle des phonèmes a été utilisé pour l'évaluation qui consistait à comparer la représentation des paramètres acoustiques dans les cas d'utilisation de STRAIGHT et de SPTK (Fukada *et al.*, 1992). Les tests MOS ont montré que l'utilisation de STRAIGHT permet d'obtenir une meilleure qualité de parole par rapport à l'utilisation du système HTS standard.

3.3.5 Synthèse par approche neuronale

L'utilisation des réseaux de neurones s'est avérée bénéfique dans de nombreux domaines, y compris pour la synthèse de la parole. (Rebai et BenAyed, 2015) ont généré des signaux de parole arabe en utilisant un système de synthèse de parole dont la modélisation acoustique fait appel à l'utilisation des réseaux de neurones. Ce même système inclut un système de diacritisation basé sur les réseaux de neurones (figure 3.1).

Généralement les textes arabes ne contiennent pas de signes diacritiques (les voyelles courtes), ce qui rend difficile la connaissance de la prononciation exacte du mot. Le système de diacritisation a pour objectif la restauration des signes diacritiques qui ne sont pas indiqués dans le texte. Il est basé sur un modèle génératif : les réseaux de neurones dans ce cas.

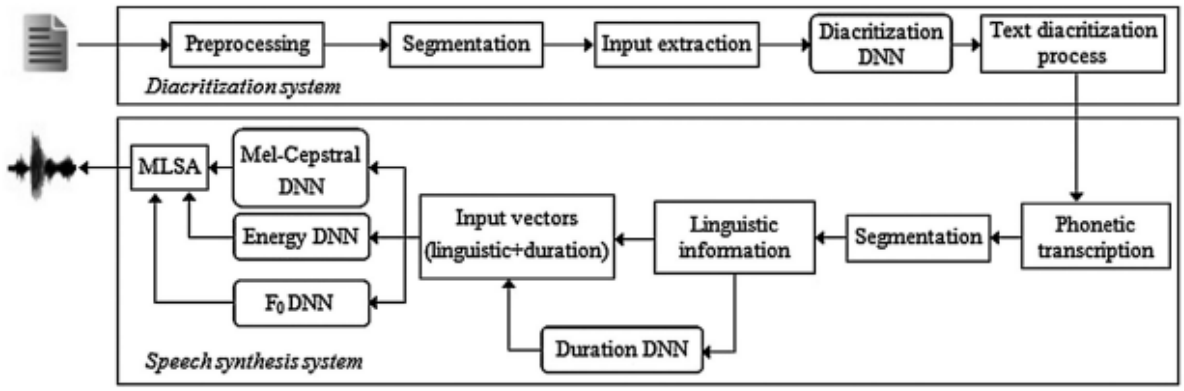


FIGURE 3.1 – Système TTS avec diacritisation du texte. [2]

Un modèle est appris en utilisant un texte avec des signes diacritique pour prédire ceux du texte non voyellé. Les signaux générés en utilisant le système décrit dans la figure 3.1 ont été évalués par un ensemble des tests d'écoute afin de juger leur qualité et leur intelligibilité. Les résultats ont montré que la parole synthétisée a plutôt une bonne qualité et est intelligible (des score MOS de 4 sur une échelle de 1 (très mauvaise) à 5 (excellente)).

D'autres travaux se sont intéressés à l'utilisation des réseaux de neurones avec l'algorithme de synthèse TD-PSOLA, tel est le cas de (Zaki *et al.*, 2001). Les réseaux de neurones ont été utilisés afin de générer le contour de la fréquence fondamentale. Le choix de cette architecture a été justifié par le fait qu'elle associe directement la description linguistique à la prosodie. Cette description a été faite à l'échelle de la syllabe, ainsi à chaque unité syllabique, correspond un ensemble d'informations phonétiques, phonologiques et prosodiques qui constituent les paramètres du vecteur d'entré du réseau de neurones. Le vecteur de sortie contient deux valeurs de F0 attribuées à chaque phonème de la syllabe : chaque phonème possède une valeur de F0 de début et de fin. En pratique, une technique de fenêtrage a été utilisée pour prédire le contour de F0, une méthode inspirée du système NETalk (Sejnowski, 1987). Pour chaque entrée, la fenêtre inclut les informations linguistiques de la syllabe courante ainsi que celles des syllabes précédente et suivante. Un corpus de taille réduite a été utilisé pour l'évaluation du système de synthèse proposé, ce qui explique la dégradation de la qualité des signaux générés par rapport aux signaux de parole naturelle. Cependant, le contour de F0 généré la même forme que celui de la parole naturelle.

2. Source : (Rebai et BenAyed, 2015)

3.3.6 Synthèse de la parole arabe expressive

Certains travaux se sont intéressés à la synthèse de la parole arabe expressive. La méthode décrite dans (Al-Dakkak *et al.*, 2005), permet de faire la synthèse de phrases en arabe en introduisant cinq émotions : joie, peur, colère, tristesse et surprise. Pour ce faire, une approche de synthèse par règles a été combinée avec le synthétiseur MBROLA (Dutoit et Pagel, 1996). (Azmy *et al.*, 2013) propose un système de génération de trois styles expressifs à savoir le style normal, triste et question en se basant sur la méthode de synthèse par sélection d'unités.

(Demri, 2016) s'est intéressé tout d'abord à la conception d'un corpus de diphones et d'émotions afin de synthétiser des signaux de parole expressive. L'approche adoptée a été la concaténation de diphones de styles expressifs différents (neutre, joie, tristesse et colère) pour la génération des diverses expressivités. Le corpus des expressions a été évalué à travers un site web avec la participation d'auditeurs non-experts. Les résultats des évaluations ont montré que les styles expressifs générés sont bien reconnus par les auditeurs.

3.4 La modélisation des unités de la parole

Après l'étude des caractéristiques phonétiques et phonologiques de l'arabe standard pour la synthèse paramétrique de la parole arabe, nous avons proposé un ensemble de descripteurs contextuels en se basant sur l'ensemble des descripteurs standards pour l'anglais. Notre étude a porté sur la manière de modéliser les deux classes des consonnes (simples et géminées) ainsi que les deux classes des voyelles (courtes et longues) : faut-il utiliser deux modèles pour les consonnes (un pour les consonnes géminées et un autre pour les consonnes simples) ou un seul (regroupant consonnes géminées et consonnes simples) et faut-il utiliser deux modèles pour les voyelles (un pour les voyelles courtes et un autre pour les voyelles longues) ou un seul (regroupant voyelles courtes et longues). Nous avons abouti à la proposition de quatre approches de modélisation :

— C2V2 : Dans cette approche, une consonne simple (/b/) et la consonne géminée correspondante (/bb/) sont modélisées par des unités différentes. De même une voyelle courte (/a/) et une voyelle longue (/a :/) sont représentées par des unités différentes.

— C2V1 : Dans cette deuxième approche, une voyelle courte (/a/) et la voyelle longue correspondante (/a :/) sont modélisées avec la même unité, en revanche le modèle d'une consonne géminée (/bb/) est différent de celui de la consonne simple correspondante (/b/).

— C1V2 : Dans cette troisième approche, une voyelle longue (/a :/) et une voyelle courte (/a/) sont modélisées par deux unités différentes. Cependant une consonne simple (/b/) et la consonne géminée correspondante (/bb/) sont modélisées par la même unité.

— C1V1 : Dans cette quatrième approche, une consonne géminée (/bb/) et la consonne simple correspondante (/b/) sont représentées par la même unité. De même, une voyelle courte (/a/) et la voyelle longue correspondante (/a :/) sont modélisées avec la même unité.

Dans tous les cas, les informations indiquant la nature du segment (consonne simple ou géminée d'une part, et voyelle courte ou longue d'autre part) sont introduites dans l'ensemble des descripteurs contextuels.

3.5 Descripteurs contextuels de la langue arabe

L'ensemble des descripteurs proposé (c.f Annexe B) pour appliquer l'approche de synthèse paramétrique à la langue arabe a été inspiré de l'ensemble standard des descripteurs (Tokuda *et al.*, 2002) (cf. Annexe A). L'unité de parole utilisée est le phonème.

3.5.1 A l'échelle du phonème

À ce niveau, les descripteurs standard relatifs au contexte du phonème courant et l'ensemble des positions relatives dans la syllabe à laquelle il appartient, ont été conservés :

— étiquette du phonème courant

— étiquettes des deux phonèmes précédents et des deux suivants.

— position du phonème courant dans la syllabe courante à partir du début et à partir de la fin de la syllabe.

Aucun descripteur à propos de l'aspect emphatique n'a été considéré au niveau des phonèmes. Ceci parce que le texte du corpus était transcrit de manière à distinguer

les consonnes de nature emphatiques des autres consonnes. Afin de tenir compte des particularités de la langue arabe, plus particulièrement les phénomènes phonologiques à savoir les voyelles longues et les consonnes géminées, nous avons ajoutés deux descripteurs supplémentaires :

- Est ce que le phonème courant est une consonne simple ou une consonne géminée ou n'est pas une consonne.
- Est ce que le phonème courant est une voyelle courte ou une voyelle longue ou n'est pas une voyelle.

3.5.2 A l'échelle de syllabe

A l'échelle de syllabe, moins de descripteurs (par rapport à l'ensemble standard) ont été utilisés (cf. Annexe B). Seuls les descripteurs contextuels suivants ont été gardés :

- Nombre de phonèmes dans la syllabe courante.
- Nombre de phonèmes dans la syllabe précédente.
- Nombre de phonèmes dans la syllabe suivante.
- Est ce que la syllabe courante porte un accent lexical (accent lexical) ?
- Est ce que la syllabe précédente porte un accent lexical (accent lexical).
- Est ce que la syllabe suivante porte un accent lexical (accent lexical).
- Position de la syllabe courante dans le mot (à partir du début et à partir de la fin).
- Position de la syllabe courante dans la phrase courante (à partir du début et à partir de la fin).
- Nombre de syllabes accentuées avant la syllabe courante dans la phrase courante.
- Nombre de syllabes accentuées après la syllabe courante dans la phrase courante.
- Nombre de syllabes à partir de la dernière syllabe accentuée jusqu'à la syllabe courante.
- Nombre de syllabes à partir de la syllabe courante jusqu'à la prochaine syllabe accentuée.
- L'identité de la voyelle de la syllabe courante.

La position de l'accent lexical a été prédite selon des règles adaptées de celles proposées par (Halpern *et al.*, 2009) et (Kouloughli, 2007) :

- Si le mot contient une syllabe super-lourde (qui ne peut exister qu'à la fin du mot), l'accent est porté par cette syllabe.
- Si le mot est monosyllabique, l'accent est porté par cette syllabe.
- Si le mot contient deux syllabes, l'accent est porté par la première syllabe.
- Si le mot contient trois syllabes ou plus, l'accent est porté par la pénultième syllabe.

3.5.3 A l'échelle du mot

Un premier descripteur indiquant la classe grammaticale du mot a été utilisé. Cette information marque la nature du mot : un verbe, un nom, un article etc.... Un deuxième descripteur a été utilisé afin de classifier les mots en deux ensembles : mots lexicaux (les verbes, les adjectifs, les adverbes...) et mots grammaticaux (pronoms, conjonctions...). Toujours à ce niveau, les descripteurs indiquant le nombre de syllabes dans le mot courant, précédent et suivant ainsi que la position du mot courant dans la phrase ont été utilisés :

- Tag grammatical estimé du mot précédent.
- Nombre de syllabes dans le mot précédent.
- Tag grammatical estimé du mot courant
- Nombre de syllabes dans le mot courant.
- Position du mot courant dans la phrase courante (à partir du début).
- Position du mot courant dans la phrase courante (à partir de la fin).
- Nombre de mots grammaticaux/lexicaux avant le mot courant dans la phrase courante.
- Nombre de mots grammaticaux/lexicaux après le mot courant dans la phrase courante.
- Tag grammatical estimé du mot suivant.
- Nombre de syllabes dans le mot suivant.

3.5.4 A l'échelle de phrase et de l'énoncé

Au niveau de la phrase, les mêmes descripteurs que ceux de l'ensemble standard ont été utilisés sauf celui concernant l'information ToBI car jusqu'à maintenant, aucun système ToBI n'a été développé pour la langue arabe (Kouloughli, 2007), (Al-Ani, 1970) :

- Nombre de syllabes dans la phrase précédente.
- Nombre de mots dans la phrase précédente.
- Nombre de syllabes dans la phrase courante.
- Nombre de mots dans la phrase courante.
- Position de la phrase dans l'énoncé (à partir du début).
- Position de la phrase dans l'énoncé (à partir de la fin).
- Nombre de syllabes dans la phrase suivante.
- Nombre de mots dans la phrase suivante.
- Nombre de syllabes dans l'énoncé.
- Nombre de mots dans l'énoncé.
- Nombre de phrases dans l'énoncé.

3.6 Données expérimentales

Cette partie est consacrée à la description des données utilisées dans la partie expérimentale de nos travaux, ainsi que les particularités du corpus.

3.6.1 Description du corpus

Le corpus utilisé pour les travaux de cette thèse pour l'apprentissage et l'évaluation est tiré des travaux de thèse de (Halabi, 2015). Le corpus contient 1806 phrases enregistrées en arabe Standard. Elles sont prononcées par un locuteur homme d'origine levantine (pays de Moyen-orient). Le corpus est d'une durée de 3h45 heures. Le texte des enregistrements est annoté avec une transcription orthographique (en utilisant le codage de Buckwalter) et une transcription phonétique. Les phrases prononcées par le locuteur sont tirées des bulletins d'informations de types sportifs, politiques ou météorologiques. Afin de garantir une variété des contextes, des phrases sans sens ont été enregistrées et ajoutées au corpus. Elles sont toutes prononcées dans un style neutre, donc sans exprimer des émotions. Les signaux audios ont été enregistrés dans un studio professionnel et échantillonnés à 48 kHz.

Le corpus a été conçu pour être utilisé principalement pour la synthèse de la parole arabe à partir du texte. Il a été évalué avec le système de synthèse de parole décrit dans (Halabi, 2015).

3.6.2 Analyse et traitement des données expérimentales

Nous avons fait des statistiques sur le corpus. Les résultats montrent qu'il couvre les différents types de consonnes que ce soit en termes de classes (plosives, fricatives...) ou de modes d'articulation (glottal, bilabial, dental...). Le corpus contient des mots étrangers, ce qui nécessité une transcription de certaines lettres qui n'existe pas en arabe, telle que /p/ et /v/. La figure 3.2 donne le nombre d'occurrences de phonèmes pour les quatre classes considérées.

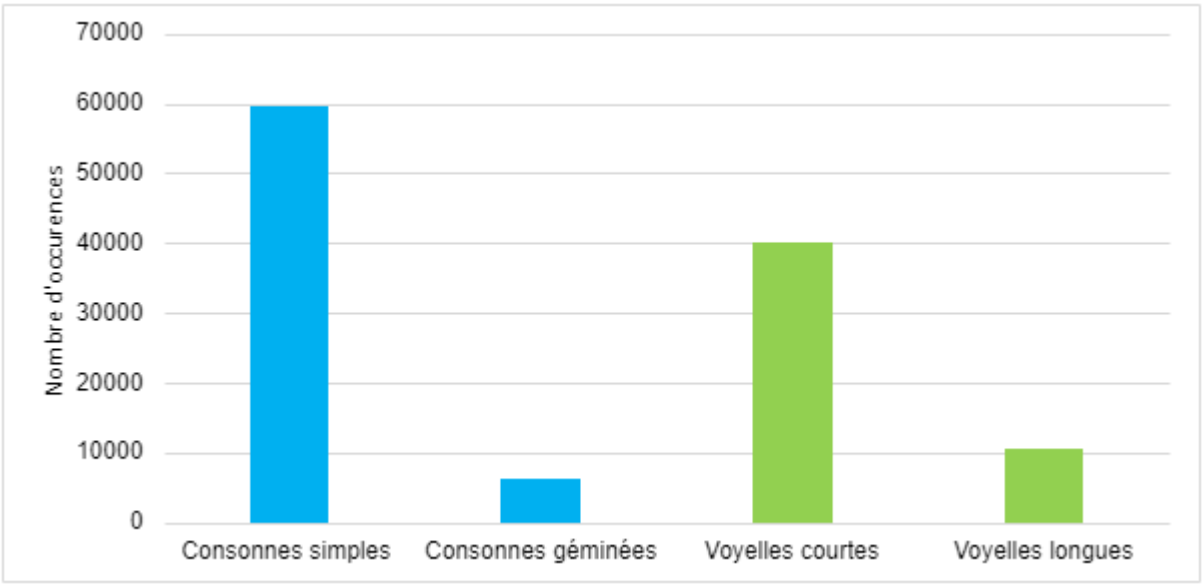


FIGURE 3.2 – Nombre d'occurrences par classe de phonèmes

La figure 3.3 indique la distribution des durées des phonèmes pour chacune des classes. Les résultats obtenus montrent que la durée originale d'une voyelle courte (respectivement consonne simple) est la moitié de la durée originale de la voyelle longue correspondante (respectivement consonne géminée).

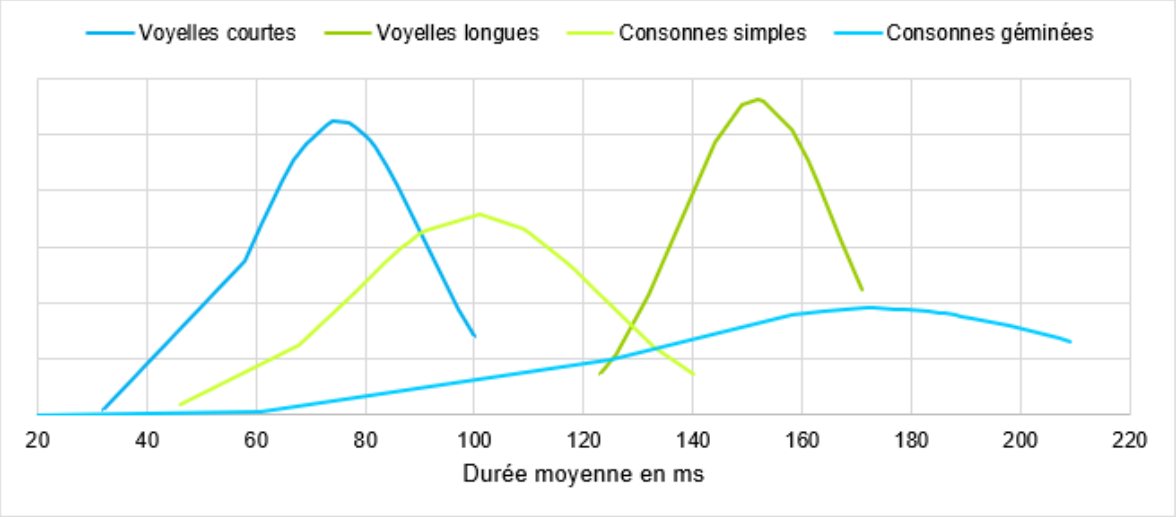


FIGURE 3.3 – Durée moyenne selon les classes de phonèmes

L'annotation du corpus tient compte des phénomènes phonologiques de la langue arabe, tels que la gémination des consonnes et les consonnes à aspect emphatique. Les transcriptions phonétiques du corpus utilisent des symboles pour distinguer les particularités phonologiques de la langue arabe

Pour la suite des expériences, l'utilisation des deux toolkit HTS et MERLIN a nécessité l'adaptation de la transcriptions phonétiques aux contraintes exigés ; plus particulièrement le choix de codage des phonèmes (Table 3.1). La segmentation du corpus a été faite d'une manière automatique et avec quelques corrections à faire manuellement. De ce fait, nous avons eu besoin d'autres structures syllabiques (à part celles mentionnées plus haut) pour réaliser la syllabification du texte telles que la structure "CCVCC". Ces nouvelles structures ont été introduites afin de pouvoir appliquer les règles de prédiction de la position de l'accent lexical dans les mots.

TABLE 3.1 – Transcription des phonèmes arabes

Lettre	Phonème IPA	Notation Buckwalter	HTS et MERLIN
أ	/ʔ/	>	e
ب	/b/	B	b
ت	/t/	T	t
ث	/θ/	^	c
ج	/ʤ/	J	j
ح	/ħ/	H	H
خ	/x/	X	kh
د	/d/	d	d
ذ	/ð/	*	th
ر	/r/	r	r
ز	/z/	z	z
س	/s/	s	s
ش	/ʃ/	$	ch
ص	/sˤ/	S	S
ض	/dˤ/	D	D
ط	/tˤ/	T	T
ظ	/ðˤ/	Z	Z
ع	/ʕ/	E	E
غ	/ɣ/	G	G
ف	/f/	f	f
ق	/q/	q	q
ك	/k/	k	k
ل	/l/	l	l
م	/m/	m	m
ن	/n/	n	n
ه	/h/	h	h
و	/w/	w	w
ي	/aj/	y	y

3.7 Conclusion

Ce chapitre a présenté en premier lieu, la langue arabe et ses particularités phonologiques. Cela a été nécessaire pour adapter à la langue arabe l'ensemble des descripteurs contextuels nécessaires pour la synthèse paramétrique. Puis nous avons présenté les travaux antérieurs faits sur la synthèse de parole de la langue arabe en utilisant les différentes techniques telles que la synthèse par règles, par concaténation ou encore la synthèse paramétrique basée sur les HMM.

La deuxième partie du chapitre a été consacrée à nos travaux de thèse. Cela a constitué la phase préliminaire pour l'application des différentes approches paramétriques (par HMM et par DNN) de synthèse de la parole arabe qui seront présentées dans les chapitres qui suivent. Nous avons en particulier présenté la problématique de modélisation des unités phonétiques, ainsi que l'adaptation des descripteurs contextuels pour la synthèse paramétrique de l'arabe.

Chapitre 4

Synthèse de la parole arabe par HMM

Sommaire

4.1 Introduction

Une première contribution de cette thèse concernait la synthèse de parole arabe par HMM. Les expériences ont été menées avec HTS, après extraction des descripteurs contextuels correspondants à la langue arabe et à ses particularités. Les signaux synthétisés ont été évalués tout d'abord objectivement en analysant les durées prédites. Ensuite, des évaluations perceptives ont été conduites. L'évaluation concernait à la fois la synthèse par les HMM et la comparaison des quatre approches de modélisation des unités de parole.

Les travaux décrits dans ce chapitre ont été présentés dans les articles (Houidhek *et al.*, 2017) et (Houidhek *et al.*, 2018b).

4.2 Synthèse de la parole arabe par HMM

4.2.1 Rappel du principe de synthèse de parole par HMM

Il s'agit d'une approche paramétrique statistique : elle utilise un ensemble des paramètres pour décrire le signal de parole et des statistiques pour décrire les paramètres. L'utilisation des HMM dans la synthèse de parole est basée sur l'emploi de HTK (Young et Young, 1993). Le système résultant est appelé HTS (Black *et al.*, 2007). La synthèse de parole par HTS présente certains avantages tels que le changement des caractéristiques des signaux de parole à synthétiser, et l'utilisation d'une mémoire réduite par rapport à l'approche par sélection d'unités. La performance de la synthèse de parole par les HMM (Hidden Markov Model) est étroitement liée à la paramétrisation du signal de parole ainsi qu'à la modélisation des unités de parole. Pendant la mise en oeuvre, le système HTS utilise des modèles des phonèmes dépendants du contexte ce qui exige une description de chaque unité de parole par un ensemble des descripteurs contextuels qui contient les facteurs affectants, ou pouvant affecter la prononciation du phonème (Tokuda *et al.*, 2002).

Le mécanisme de synthèse passe par deux phases principales (Black *et al.*, 2007). L'étape d'apprentissage commence d'abord par appliquer les méthodes de traitement du texte, à savoir la normalisation, la segmentation et l'annotation. Ensuite, l'extraction des paramètres acoustiques des signaux de parole. L'ensemble des paramètres comporte les paramètres spectraux et leurs dérivées premières et secondes, les paramètres d'excitation tels que la fréquence fondamentale et les paramètres d'apériodicité ((Kawahara *et al.*, 2001)). Le grand nombre des descripteurs contextuels utilisées pour qualifier le contexte des unités de parole rend plus compliqué le choix des contextes essentiels et utiles, c'est ainsi que les arbres de décisions sont introduits afin de regrouper les modèles ayants des fonctions des densité de probabilités similaires. Les densités similaires partagent alors les mêmes paramètres. À la fin de l'apprentissage, trois modèles HMM de prédiction sont obtenus : un pour la durée, un autre pour les paramètres spectraux et un troisième pour la fréquence fondamentale.

L'étape de synthèse commence tout d'abord par convertir le texte à synthétiser en

une séquence des descripteurs contextuels. La phrase HMM correspondante est formée par concaténation des HMM dépendants du contexte correspondant aux unités de parole formant la phrase. Les durées des états de l'énoncé HMM sont déterminées afin de maximiser la probabilité de sortie. Les paramètres spectraux et ceux d'excitation sont générés à partir des HMM à l'aide d'un algorithme de génération de paramètres, MLPG (Maximum Likelihood Parameter Generation) qui optimise les probabilités de sortie.

Finalement, le filtre de synthèse MLSA (Mel-Log Spectrum Approximation), produit le signal de parole en utilisant les paramètres d'excitation et de spectre générés. À ce niveau, il est important d'indiquer que le choix des descripteurs contextuels est très important car ils interviennent pendant l'apprentissage lors de la construction des arbres de décision pour le partage de paramètres entre les modèles HMM dépendants du contexte et dans la partie synthèse quand les HMM sont utilisés pour prédire les paramètres de parole. Ainsi, ils ont un impact considérable sur la qualité de parole générée. Une partie des descripteurs dépend de la langue, ainsi, donc pour appliquer HTS, il est nécessaire de tenir compte des particularités phonologiques et linguistiques de la langue à traiter (Zen *et al.*, 2007).

4.2.2 Adaptation de la synthèse par HMM à l'arabe

Une première étape consiste à adapter l'ensemble des descripteurs contextuels à la langue arabe, ceci fait appel aux particularités phonologiques et linguistiques de l'arabe.

Particularités de l'arabe

La langue arabe présente certaines particularités qui doivent être prises en compte lors de l'application de l'approche de synthèse paramétrique (Al-Ani, 1970). Coté phonologie, deux phénomènes sont mis en relief : la gémination et les voyelles longues.

Les voyelles longues : La langue arabe présente deux types de voyelles : courtes et longues (Newman, 2002). Acoustiquement, une voyelle longue est deux fois plus longue que la voyelle courte correspondante (Khouja et Zrigui, 2005). En plus, si une voyelle courte est remplacée par une voyelle longue, le sens du mot change. *Exemple* : "هتف" /hatafa/ qui veut dire "il a crié", avec une première voyelle courte /a/. Si elle est remplacée par la voyelle longue correspondante /a :/, le mot devient "هَاتف" /ha :tafa/ qui veut dire "il a téléphoné".

Consonnes géminées : Les consonnes de la langue arabe peuvent exister sous deux formes : simple et géminée. Acoustiquement, une consonne géminée possède une durée supérieure (double) à celle de son homologue simple (Khouja et Zrigui, 2005). Le remplacement d'une consonne simple par la consonne géminée correspondante change le sens du mot. *Exemple :* درس "darasa" qui veut dire "il a étudié", devient دَرَّسَ "darrasa" qui veut dire "il a enseigné".

Aspect emphatiques : Il s'agit de la pharyngalisation de certaines consonnes dans certains contextes (Halabi, 2015). Certaines voyelles peuvent présenter l'aspect emphatique si elles sont précédées ou suivies par une consonne emphatique (Watson, 2002). Les consonnes en arabe peuvent être classées en trois classes :

- Consonnes de nature emphatique : /s^{ς}/ ص; /t^{ς}/ ط; /q/ ق; /d^{ς}/ ض; /$ð^{\varsigma}$/ ظ/; /ɣ/ غ; /x/ خ.
- Consonnes qui peuvent être emphatiques dans certains contextes : /r/ ر; /l/ ل.
- Les autres consonnes qui ne peuvent pas être emphatiques.

Adaptation de l'approche à l'arabe

L'ensemble standard des descripteurs décrit dans (Tokuda *et al.*, 2002) a été utilisé. Les modifications apportées consistent essentiellement en l'ajout de deux descripteurs précisant la nature du segment de parole (Houidhek *et al.*, 2018b). Le premier se réfère aux consonnes, les valeurs possibles sont :

— le segment courant est une consonne simple.

— le segment courant est une consonne géminée.

— le segment courant n'est pas une consonne.

Le deuxième est lié aux voyelles, les valeurs possibles sont :

— le segment courant est une voyelle courte.

— le segment courant est une voyelle longue.

— le segment courant n'est pas une voyelle.

Il est à noter qu'aucun descripteur contextuel sur l'aspect emphatique n'a été ajoutée à l'ensemble des descripteurs contextuels car les consonnes emphatiques étaient déjà distinguées des autres consonnes dans la transcription du corpus utilisé.

La position de l'accent lexical a été prédite selon les règles proposées par (Kouloughli, 2007) et (Al-Ani, 1970). Cependant, jusqu'au moment de l'écriture de ce manuscrit, aucun système ToBI n'a été développé pour la langue arabe, ainsi le descripteur correspondant a été ignoré.

4.3 Expériences avec HTS

Afin de produire des signaux de parole en arabe en utilisant HTS, le corpus décrit dans la section 3.6 a été utilisé. Plus précisément 1565 signaux ont été dédiés à l'apprentissage du système et 30 phrases ont été gardées pour l'évaluation. Les phrases de l'ensemble d'apprentissage ont été converties en des fichiers labels dont chacun qualifie les phonèmes de chaque phrase par l'ensemble des descripteurs contextuels décrits auparavant. Le vocodeur STRAIGHT a été introduit afin de tenir compte de l'aspect apériodique de la parole et synthétiser des signaux de meilleure qualité.

Un ensemble de réglages est à faire avant d'utiliser HTS. Par exemple, la plage fréquentielle dépend du genre du locuteur. Dans le cas présent, pour un locuteur mâle, F0 varie entre 70 et 250 Hz. La même procédure a été suivie pour chacun des systèmes de modélisation qui différencient ou fusionnent les unités des consonnes géminées (resp des voyelles longues) : C1V1, C1V2, C2V1 et C2V2 (c.f section 3.4).

4.4 Évaluation objective de la durée des phonèmes

Le but de cette évaluation objective est d'estimer la performance de HTS en ce qui concerne la prédiction des durées des phonèmes pour chaque approche de modélisation. Pour chacune de ces dernières, la moyenne sur l'ensemble des voyelles des rapports entre la durée moyenne des voyelles longues (VL) et la durée moyenne des voyelles courtes équivalentes (VC) est calculée, ainsi que le rapport entre les durées moyennes des consonnes géminées (CG) et celles des consonnes simples (CS). Seuls les phonèmes avec plus de 10 occurrences pour chaque classe (consonnes simples / géminées et voyelles courtes / longues) sont considérés. Les rapports calculés à partir des durées prédites sont comparés à ceux obtenus pour les 30 phrases de test à partir de la segmentation fournie avec le corpus. Les résultats indiqués dans la table 4.1 montrent que les rapports calculés pour les quatre approches de modélisation sont similaires à ceux calculés pour la parole naturelle.

TABLE 4.1 – Rapport des durées

	VL / VC	CG / SC
Nombre d'occurences	262 / 884	104 / 1315
$C1V1$	1,7	2,1
$C1V2$	1,7	2,1
$C2V1$	1,7	2,1
$C2V2$	1,8	2,2
Originale	2,0	2,1

Afin de pouvoir identifier les raisons de cette similarité, l'arbre de décision associé aux durées, qui permet de prédire les durées des états, pour chaque modèle a été analysé. Il est à noter, que pour chaque approche de modélisation (C1V1, C1V2, C2V1 et C2V2), il existe un seul arbre de décision pour le modèle des durées des phonèmes. La figure 4.1 représente la partie supérieure de l'arbre de décision des durées du modèle C2V2.

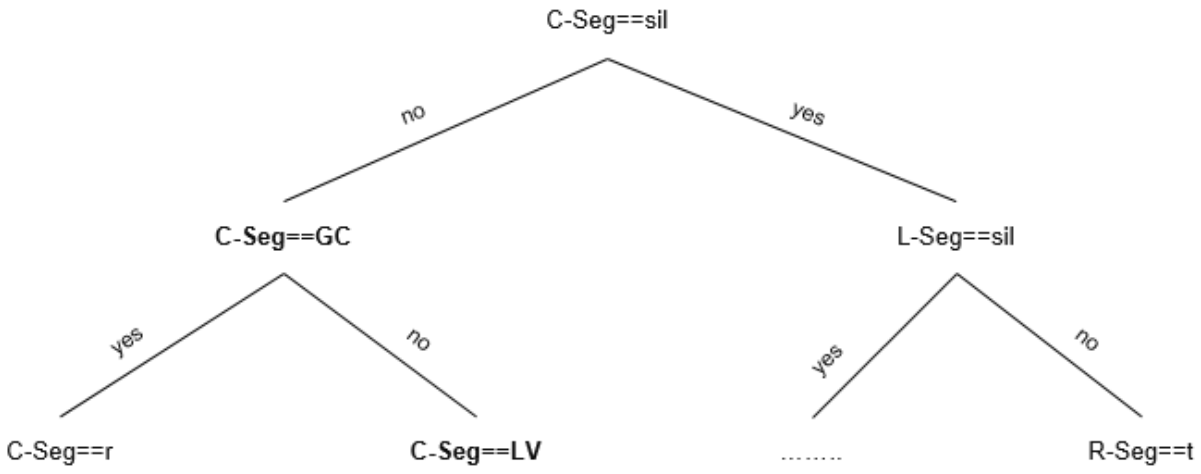


FIGURE 4.1 – Arbre de décision (partie supérieure) du modèle C2V2

Dans la figure 4.1, "C-Seg" fait référence au segment courant, "L-Seg" au segment précédent (à gauche) et "R-Seg" au segment suivant (à droite) ; "sil" indique un silence, "r" et "t" sont les phonèmes / r / et / t /, "GC" indique qu'il s'agit d'une consonne géminée et "LV" d'une voyelle longue. La figure 4.1 montre que les questions sur la nature du segment de la parole telles que consonne géminée (C-Seg == GC) et voyelle longue

(C-Seg == LV) sont situées en haut de l'arbre. Ce comportement a été observé pour les quatre modèles.

Afin de raffiner l'analyse de la similarité des rapports des durées, pour chaque modèle (C1V1, C1V2, C2V1 et C2V2), le RMSE (Root Mean Sqaure Error) entre les durées originales et les durées prédites a été calculé pour chaque classe de phonèmes (i.e., consonnes simples, consonnes géminées, voyelles courtes, voyelles longues).

Les résultats obtenus (représentés dans la figure 4.2), montrent que la valeur RMSE mesurée varie considérablement en fonction de la classe des phonèmes. Cependant, pour chaque classe, les valeurs de RMSE sont similaires et varient légèrement entre les modèles.

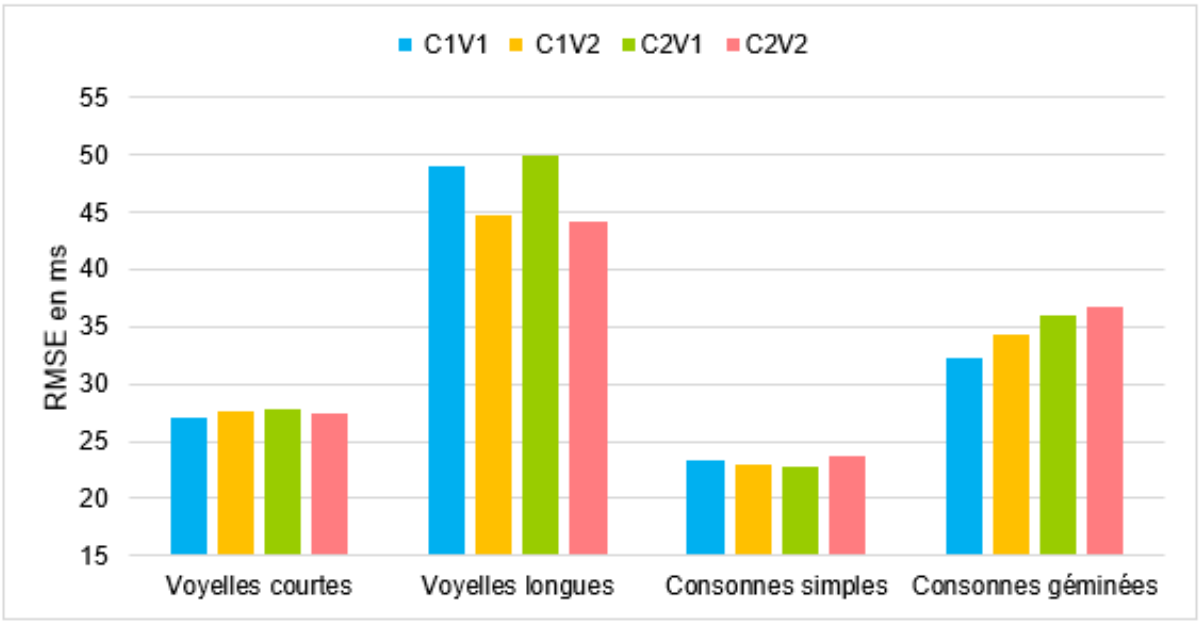


FIGURE 4.2 – RMSE entre les durées prédites et les durées naturelles

De même, le NRMSE (Normalized Root Mean Square Error) a été calculé pour chaque modèle et chaque classe des phonèmes : NRMSE = RMSE/Durée moyenne. Les résultats obtenus sont représentés dans la figure 4.3. Les résultats obtenus montrent que les NRMSE obtenus sont similaires pour les quatre modèles, entre 22% et 24% pour les consonnes géminées et entre 30% et 34% pour les voyelles courtes et longues et les consonnes simples.

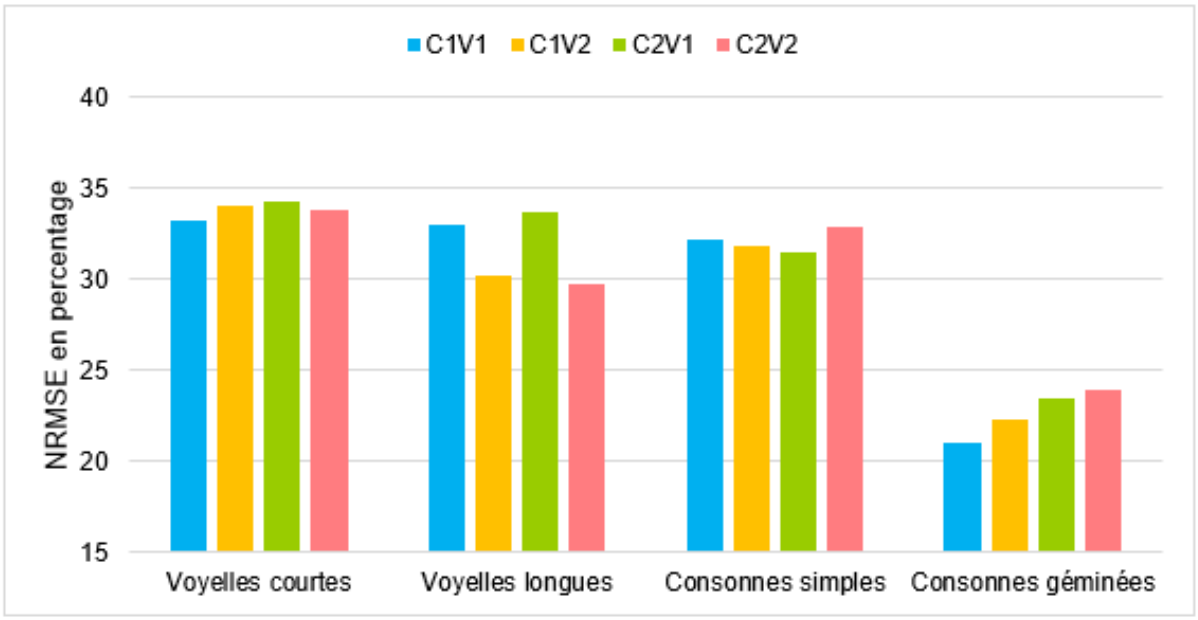


FIGURE 4.3 – NRMSE entre les durées prédites et les durées naturelles

4.5 Évaluation subjective

Des évaluations perceptives ont été conduites afin de compléter les résultats obtenus à la suite des évaluations objectives. (Wester *et al.*, 2015) a indiqué qu'il faut tout d'abord préciser les critères à évaluer. Généralement, l'évaluation des systèmes de synthèse de parole met l'accent sur l'aspect naturel et la qualité globale des signaux de parole synthétisée. De même, il nous a semblé intéressant d'évaluer le degré de dégradation causé par le système de synthèse. Les instructions des évaluations perceptives conduites sont décrites dans l'annexes C.

4.5.1 Évaluation de la qualité globale

Cette première évaluation, a pour objectif l'analyse de la qualité des signaux produits en utilisant HTS, ceci pour les quatre modèles C1V1, C1V2, C2V1 et C2V2. Deux facteurs contribuent à la qualité : l'aspect naturel et la qualité globale du signal. L'aspect naturel est évalué en se référant à l'intonation (si l'évolution du pitch est naturelle) et au rythme

(si la longueur des phonèmes parait naturelle). La qualité globale fait référence à la qualité du signal acoustique produit : l'impression globale en écoutant le signal de parole généré.

Données évaluées

Lors de l'évaluation, les ensembles suivants ont été considérés :

— les signaux synthétisés par HTS en utilisant le modèle C1V1.

— les signaux synthétisés par HTS en utilisant le modèle C1V2.

— les signaux synthétisés par HTS en utilisant le modèle C2V1.

— les signaux synthétisés par HTS en utilisant le modèle C2V2.

Pour chaque modèle, 30 phrases ont été générées (les mêmes phrases ont été produites par les quatre modèles). La durée des signaux est comprise entre 3s et 17s. Il est à noter que les phrases utilisées dans les tests perceptifs sont les mêmes que celles utilisées pour faire les mesures objectives et ne figurent évidemment pas dans le corpus d'apprentissage.

Protocole

Les tests perceptifs conduits sont de type MOS (ITU, 1996) ; ils fournissent des scores MOS. Des tests préliminaires ont été conduits où il y avait deux questions différentes afin d'évaluer la qualité globale et l'aspect naturel. Les résultats obtenus ont montré des résultats similaires pour les deux critères, ceci a été justifié par le fait que les participants n'étaient pas de spécialistes de la synthèse de parole, ainsi, ce n'était pas évident de pouvoir différencier la qualité globale de l'aspect naturel (cf. Annexe G). Dans la suite des tests, les participants devaient répondre à une seule question regroupant les deux critères : "Comment évaluez-vous la qualité globale et l'aspect naturel de ce que vous venez d'entendre par rapport à une parole naturelle (prononcée par un être humain) ?" Les réponses possibles sont des scores de 1 à 5 selon l'échelle suivante :

1. Très loin de la parole naturelle.
2. Loin de la parole naturelle
3. Un peu loin de la parole naturelle.
4. Proche de la parole naturelle.
5. Très proche de la parole naturelle.

Neufs auditeurs ont participé à ce test perceptif. Chaque participant a évalué un ensemble de 40 phrases, c'est-à-dire 10 phrases pour chacun des quatre modèles. En plus, il y avait un ensemble complémentaire de six phrases pour une phase d'introduction, afin de permettre à l'auditeur de s'adapter au son et à la répartition des scores (c.f Annexe D).

Résultats

Les scores MOS sont représentés dans la figure 4.4 avec leurs intervalles de confiance à 95%. Les scores montrent que les quatre approches de modélisation (C1V1, C1V2, C2V1 et C2V2) qui reposent sur la synthèse par HMM produisent des signaux de qualité. Les signaux issus des quatre modèles présentent une similarité au niveau des caractéristiques d'intonation et de rythme.

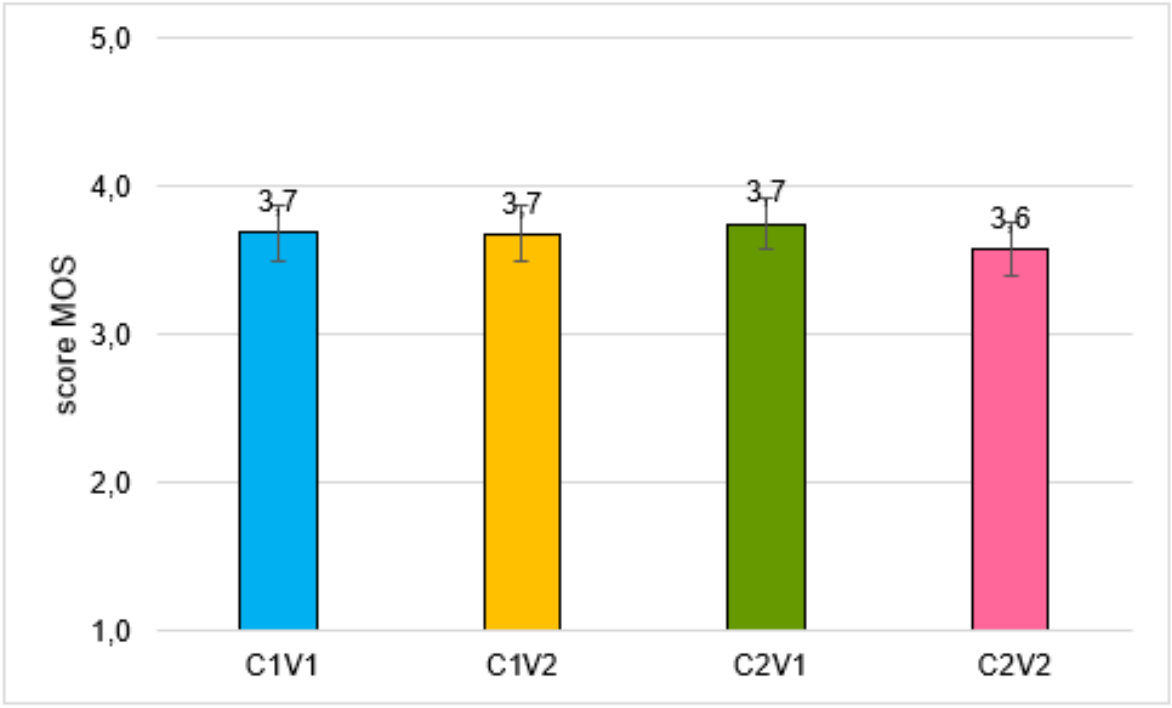


FIGURE 4.4 – Évaluation de la qualité globale et l'aspect naturel

4.5.2 Résultats de l'évaluation DMOS

Ce deuxième test consiste à évaluer le taux de dégradation des signaux générés par HTS (ITU, 1996) par rapport aux signaux naturels.

Données évaluées

Lors du test, les ensembles des signaux suivants ont été utilisés :

— les signaux synthétisés par HTS en utilisant le modèle C1V1.

— les signaux synthétisés par HTS en utilisant le modèle C1V2.

— les signaux synthétisés par HTS en utilisant le modèle C2V1.

— les signaux synthétisés par HTS en utilisant le modèle C2V2.

— les signaux naturels.

Il s'agit d'un test différentiel qui consiste en une comparaison des signaux générés par HTS aux signaux naturels correspondants (signaux de référence). Pendant l'évaluation, le signal de référence est présenté d'abord suivi, du signal synthétisé avec l'une des quatre approches de modélisation. Ce test permet d'obtenir des scores dits DMOS.

Protocole

Douze auditeurs ont participé à cette évaluation. Chacun a écouté un ensemble de 20 paires de signaux, chaque paire comprenant le signal naturel suivi par le même signal synthétisé avec l'un des quatre modèles. Les auditeurs ont noté la dégradation perçue en répondant à la question : "Comment évaluez-vous la dégradation du deuxième signal par rapport au premier ?" Les réponses possibles sont sur une échelle à cinq points de dégradation (cf. Annexe E) :

1. Dégradation très gênante
2. Dégradation gênante
3. Dégradation un peu gênante
4. Dégradation audible mais pas gênante
5. Dégradation inaudible

Résultats

Les scores DMOS obtenus sont représentés dans la figure 4.5 associés à leurs intervalles de confiance à 95%.

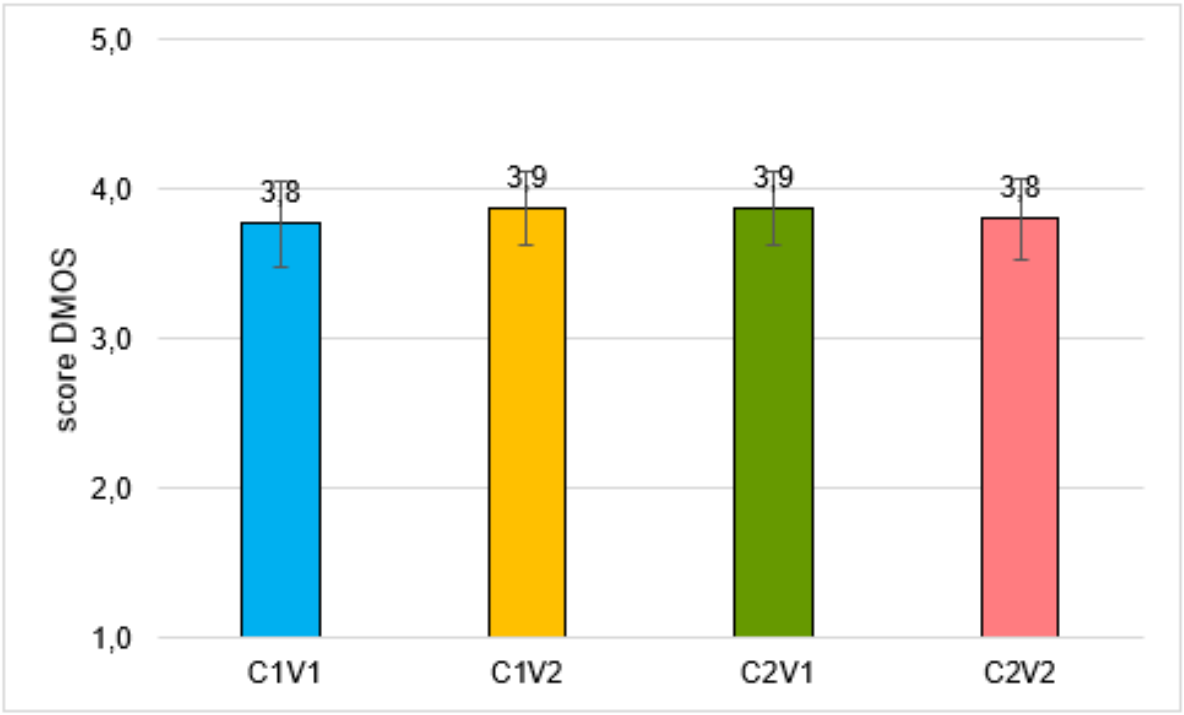


FIGURE 4.5 – Résultats de l'évaluation DMOS

Plus le score est élevé, plus la dégradation est faible. Les scores DMOS obtenus montrent que les signaux produits avec les quatre modèles C1V1, C1V2, C2V1 et C2V2 présentent un degré de dégradation similaire. Ces résultats sont cohérents avec le fait que les quatre systèmes offrent une qualité globale similaire d'après les résultats de l'évaluation de la qualité globale.

4.5.3 Comparaison des modèles

Les phrases synthétisées avec les quatre modèles (C1V1, C1V2, C2V1 et C2V2) ont été comparées deux à deux. Ceci vise à déterminer s'il y a une préférence pour une approche particulière de modélisation. Chaque auditeur écoute des paires de phrases (provenant de

modèles différents) et doit indiquer sa préférence.

Données évaluées

Pendant le test d'écoute, les phrases suivantes sont évaluées :

— les signaux synthétisés par HTS en utilisant le modèle C1V1.

— les signaux synthétisés par HTS en utilisant le modèle C1V2.

— les signaux synthétisés par HTS en utilisant le modèle C2V1.

— les signaux synthétisés par HTS en utilisant le modèle C2V2.

Pendant les expériences, les combinaisons possibles lors de la comparaison sont :

— C1V1 / C1V2

— C1V1 / C2V1

— C1V1 / C2V2

— C1V2 / C2V1

— C1V2 / C2V2

— C2V2 / C2V1

Protocole

Le test de comparaison (ITU, 1996) a été conduit afin de comparer les quatre modèles selon les combinaisons présentées précédemment. Vingt-sept locuteurs natifs arabes ont participé à cette évaluation. Chacun a évalué un ensemble de 23 paires de signaux vocaux ; chaque paire est constituée du même énoncé produit avec deux systèmes différents. L'ordre de présentation des signaux de parole est choisi au hasard pour chaque écoute d'une paire. Au cours de l'évaluation, les participants ont été invités à indiquer le signal préféré en fonction de la qualité globale de la parole produite, en répondant à la question "Comment jugez-vous la qualité du second signal par rapport au premier ?". Les réponses possibles sont des scores allant de 1 "beaucoup plus mauvaise", à 7 "Bien meilleure" (cf. Annexe F).

1. Beaucoup plus mauvaise
2. Plus mauvaise
3. Un peu plus mauvaise
4. A peu près la même

5. Un peu meilleure
6. Meilleure
7. Bien meilleure

Résultats

Afin de faciliter l'analyse des résultats, les scores ont été regroupés en trois catégories : « premier préféré », « pas de préférence » et « deuxième préféré», comme suit :

$$\left.\begin{array}{ll} 1 & \textbf{Beaucoup plus mauvaise} \\ 2 & \textbf{Plus mauvaise} \end{array}\right\} = \textbf{Premier préféré}$$

$$\left.\begin{array}{ll} 3 & \textbf{Un peu plus mauvaise} \\ 4 & \textbf{A peu près la même} \\ 5 & \textbf{Un peu meilleure} \end{array}\right\} = \textbf{Pas de préférence}$$

$$\left.\begin{array}{ll} 6 & \textbf{Meilleure} \\ 7 & \textbf{Bien meilleure} \end{array}\right\} = \textbf{Deuxième préféré}$$

Après regroupement des scores, les résultats ont été analysés et représentés dans la figure 4.6. Par exemple, la ligne inférieure correspond à la comparaison de C1V2 avec C2V1. Pour cette ligne, 12% des réponses donnent une préférence au premier modèle (côté gauche, c'est-à-dire C1V2), 6% des réponses préfèrent le second modèle (côté droit, c'est-à-dire C2V1) et la partie centrale montre que 82% des réponses expriment aucune préférence. La comparaison individuelle des quatre modèles montre que les auditeurs n'avaient aucune préférence claire pour un système particulier. Par conséquent, la différenciation des unités pour les consonnes géminées (resp. les voyelles longues) des unités associées aux consonnes simples (resp. aux voyelles courtes) ou leur fusion conduit à une qualité de synthèse de la parole similaire, dans le cas de la synthèse paramétrique de l'arabe par HMM.

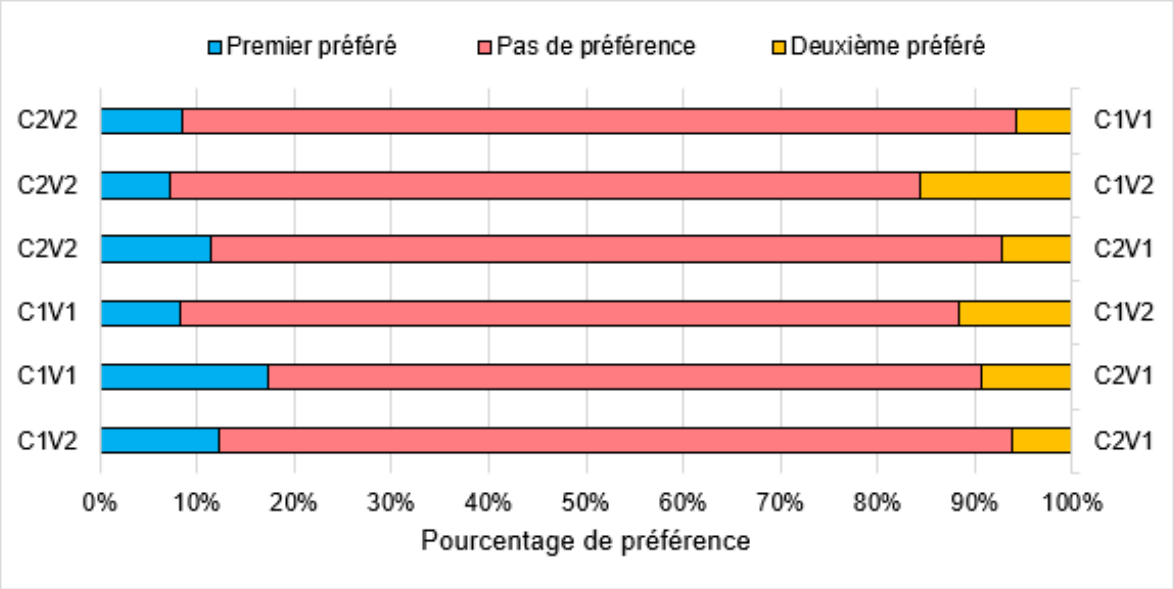


FIGURE 4.6 – Comparaison des quatre modèles

4.6 Conclusion

Dans ce chapitre, les différentes approches possibles de la modélisation des unités vocales pour la synthèse de la parole en arabe basée sur les HMM (HTS) ont été évaluées. Les mesures objectives montrent que les durées des segments générées avec HTS pour les quatre modèles sont similaires. Cette conclusion a été confirmée par des tests d'écoute (MOS, DMOS et tests de préférences). Les résultats ont montré qu'il n'y avait pas de différence importante entre les différentes approches de modélisation. Ainsi, l'identification des consonnes géminées et des voyelles longues en tant que phonèmes à part entière (donc modélisés par des unités spécifiques) n'est pas obligatoire lors de l'application de HTS pour la synthèse de la parole arabe, tant que ces informations existent dans l'ensemble des caractéristiques contextuelles.

Le chapitre suivant traite la synthèse de la parole en arabe basée sur des approches d'apprentissage profond (Zen *et al.*, 2013). Récemment, différentes variantes et architectures des réseaux de neurones profonds (DNN) ont été introduites pour la synthèse de la parole (Zen et Sak, 2015) (Wu *et al.*, 2016) et les résultats obtenus ont montré que l'utilisation

des DNN améliore la qualité et le naturel des signaux. Ainsi, il est intéressant d'analyser si l'approche DNN bénéficie de la différenciation explicite des consonnes géminées par rapport aux consonnes simples et des voyelles longues par rapport aux voyelles courtes, contrairement à ce qui a été observé avec l'approche de synthèse par HMM.

Chapitre 5

Synthèse de la parole arabe par DNN

Sommaire

5.1 Introduction

La qualité des signaux de parole générés par HTS n'a pas encore atteint le même niveau que celle d'un système de synthèse par sélection d'unités. En outre, les expériences décrites

dans le chapitre 4, ont montré que les HMM n'ont pas profité de la différentiation des consonnes simples / consonnes géminées (resp. voyelles longues / voyelles courtes). (Zen, 2013) et (Fan *et al.*, 2014) ont montré que l'utilisation des DNN dans la synthèse permet d'améliorer la qualité de la parole générée, il est ainsi intéressant d'investiguer si les DNN traitent différemment la modélisation des unités de parole.

Dans ce chapitre, les expériences de synthèse de parole arabe basée sur l'utilisation des DNN sont décrites. Le toolkit MERLIN a été utilisé tout au long de cette partie expérimentale. La mise en œuvre de MERLIN nécessite de choisir un type d'alignement (par phonème ou par état) ainsi que l'architecture. Après la génération des signaux de parole, des évaluations objectives et subjectives ont été conduites afin d'estimer les performances des DNN en synthèse de parole arabe.

5.2 Synthèse de la parole arabe par DNN

Malgré les modifications apportées au système de synthèse par HMM, la qualité des signaux générés demeure moins bonne que le celle des signaux générés par sélection des unités (Zen *et al.*, 2013) et (Black *et al.*, 2007). Ceci est dû principalement aux trois facteurs suivants ; l'utilisation du vocodeur, le modèle acoustique imprécis, et le sur-lissage. L'utilisation de l'approche paramétrique statistique de synthèse par HMM repose sur les modèles acoustiques qui assurent le passage des descripteurs contextuels aux paramètres acoustiques. En pratique, cette correspondance est assurée par les arbres de décision (Jurafsky et James, 2000). Il s'agit d'architectures peu profondes qui sont jugées inefficaces pour représenter les dépendances entre les descripteurs contextuels et les paramètres acoustiques car elles présentent certaines défaillances (Watts *et al.*, 2016).

Les arbres de décision n'offrent pas la possibilité de modéliser des fonctions complexes similaires à celles qui existent entre les descripteurs contextuels et les paramètres acoustiques. Comme l'ensemble des descripteurs contextuels contient environ 50 informations, il est nécessaire d'estimer de grands arbres de décision. En outre, pendant l'apprentissage, les arbres de décision regroupent les données en sous-groupes et utilisent des paramètres différents pour chaque groupe (Zen, 2013). Ce processus affecte le groupement des distributions dépendantes du contexte, et par la suite l'estimation des distributions pour la prédiction des paramètres de parole. Une solution à ces problèmes est le remplacement des arbres de décision par des DNN. En effet, selon (Bengio *et al.*, 2009), les DNN sont

capables de représenter des fonctions complexes.

5.2.1 Rappel du principe de la synthèse de parole par DNN

Similairement à la synthèse par HMM, le processus de synthèse par DNN commence par convertir le texte en une séquence des descripteurs contextuels. Le même ensemble de descripteurs que celui décrit précédemment est utilisé. Le passage des descripteurs contextuels aux durées et aux paramètres acoustiques (les paramètres spectraux et d'excitation ainsi que leurs dérivées) est réalisé par des DNN. Les paramètres des DNN sont mis à jour à partir de l'ensemble des données d'apprentissage : ils sont appris à l'aide de paires (vecteurs d'entrée et de sortie extraits des données d'apprentissage) afin de minimiser l'erreur entre la sortie prédite à partir de l'entrée donnée et la sortie cible. Enfin, un vocodeur est également utilisé pour traiter les paramètres de parole générés afin de produire le signal de parole correspondant au texte. Le vecteur des entrées des DNN contient des informations numériques et des informations binaires qui sont les réponses aux questions sur le phonème et sur son contexte. Les travaux décrits dans ce chapitre ont été présentés dans l'article (Houidhek *et al.*, 2018a)

5.2.2 Adaptation de la synthèse par DNN à la langue arabe

L'utilisation du toolkit MERLIN pour la synthèse de parole a suivi le même principe que celui décrit dans le chapitre précédent (4) lors de l'adaptation de HTS pour l'arabe. L'ensemble standard des descripteurs contextuels (Tokuda *et al.*, 2002) a été adapté aux caractéristiques de la langue arabe (section 3.2). Deux descripteurs contenant les informations sur la nature du segment courant ont été ajoutés.

— Le premier est lié aux consonnes, les valeurs possibles sont :
  - le segment courant est une consonne simple.
  - le segment courant est une consonne géminée.
  - le segment courant n'est pas une consonne.

— Le deuxième est lié aux voyelles, les valeurs possibles sont :
  - le segment courant est une voyelle courte.
  - le segment courant est une voyelle longue.
  - le segment courant n'est pas une voyelle.

L'accent lexical a été positionné selon les règles proposées par (Kouloughli, 2007) et (Al-Ani, 1970). Pareillement aux expériences avec HTS, le critère ToBI n'a pas été introduit car jusqu'au moment de l'écriture de ce manuscrit, aucun système ToBI n'a été développé pour la langue arabe.

5.3 Expériences avec MERLIN

Pour les expériences, le corpus décrit dans la section 3.6 a été utilisé. La découpe des données est comme suit : 1565 signaux ont été dédiés à l'apprentissage du système et 30 phrases ont été gardées pour l'évaluation. Il s'agit de la même partition des données expérimentales que celles utilisée dans le chapitre 4. Le vocodeur WORLD (Morise *et al.*, 2016) a été associé à MERLIN. (Wu *et al.*, 2016) a montré que les deux vocodeurs STRAIGHT et WORLD (section 2.2.3) permettent de générer des signaux de qualité similaire. STRAIGHT est utilisé pour extraire les coefficients Mel-Cepstraux (MCC) de dimensions 60, 25 bandes d'apériodicités (BAP) et la fréquence fondamentale sur une échelle logarithmique (log F0) chaque 5 ms. WORLD permet d'extraire des MCC de dimensions 60, 5 bandes d'apériodicité et log F0 toutes les 5 ms. Ainsi, le vecteur de sortie des réseaux de neurones est donc constitué de MCC, BAP et log F0 avec leurs dérivées premières et dérivées secondes, à ces paramètres s'ajoute une information binaire indiquant le voisement ou le non-voisement des segments de parole.

5.3.1 Choix du type d'alignement

L'alignement est la détermination des frontières exactes des unités de parole (exemple : phonème). Ainsi, le signal de la parole sera aligné à la représentation phonétique correspondante. L'alignement peut être fait manuellement ou automatiquement. La précision de l'alignement est dépendante de la qualité des enregistrements, et affecte la qualité de la parole synthétisée. Il est possible d'utiliser soit l'alignement par phonème soit par état lors de la synthèse de parole par DNN. (Watts *et al.*, 2016) a montré que l'utilisation d'un alignement par état améliore la qualité des signaux de parole synthétisés. Un ensemble de tests a été conduit afin de confirmer l'hypothèse pour la langue arabe et lorsque l'on utilise MERLIN.

Mesures objectives avec alignements par phonème ou par état

Dans ce test, une architecture de six couches cachées Feed Forward standards a été utilisée ; chaque couche comporte 1024 nœuds et une fonction d'activation de type tangente hyperbolique. L'évaluation objective des signaux synthétisés en utilisant les deux types d'alignement (par phonème et par état) consiste en un calcul de certaines mesures objectives telle que la distorsion des coefficients cepstraux, la distorsion de l'apériodicité. Plus la valeur est faible (que ce soit pour la distorsion, l'erreur ou le voisement), meilleures sont les performances. Les valeurs obtenues sont représentées dans le tableau 5.1, elles montrent que l'alignement par état permet d'avoir une meilleure prédiction des coefficients spectraux, de l'apériodicité, de la fréquence fondamentale, du voisement et de la durée, conduisant ainsi à une meilleure qualité des signaux synthétisés comparée à celle des signaux générés avec l'alignement par phonème. Cette conclusion a été confirmée en faisant l'écoute de quelques signaux générés par un alignement par phonème et de quelques signaux générés avec un alignement par état. Une meilleure qualité a été perçue pour les signaux générés avec un alignement par état.

TABLE 5.1 – Comparaison des alignements

Mesures	**Phonème**	**État**
Distorsion des coefficients Mel-cepstraux MCD (dB)	5,3	**4,4**
Distorsion de la bande d'apériodicité BAP (dB)	0,4	**0,3**
Erreur quadratique de F0 RMSE F0 (Hz)	12,8	**11,1**
Erreur de voisement VUV (%)	7,5	**5,8**
Erreur quadratique de durée RMSE (ms)	28,5	**27,5**

5.3.2 Choix d'architecture

Avec MERLIN, il est possible d'implémenter différentes architectures pour faire la synthèse de parole à partir du texte. Il est important de choisir celle qui permet d'avoir la meilleure qualité des signaux générés. Ainsi, quelques expériences ont été menées en utilisant différentes architectures. Nous avons essayé les architectures suivantes :

— DNN (Deep Neural Network)
  — Architecture : ['TANH', 'TANH', 'TANH', 'TANH', 'TANH', 'TANH']
  — Taille des couches cachées : [1024, 1024, 1024, 1024, 1024, 1024]

- Algorithme d'optimisation : adam
- Vocodeur : WORLD

— LSTM

- Architecture : ['TANH', 'TANH', 'TANH', 'TANH', 'LSTM', 'LSTM']
- Taille des couches cachées : [1024, 1024, 1024, 1024, 512, 512]
- Algorithme d'optimisation : adam
- Vocodeur : WORLD

— GRU (Gated Recurrent Unit)

- Architecture : ['TANH', 'TANH', 'GRU', 'GRU']
- Taille des couches cachées : [512, 512, 230, 230]
- Algorithme d'optimisation : adam
- Vocodeur : WORLD

— BLSTM

- Architecture : ['TANH', 'TANH', 'TANH', 'TANH', 'BLSTM']
- Taille des couches cachées : [1024, 1024, 1024, 1024, 512]
- Algorithme d'optimisation : adam
- Vocodeur : WORLD

Où "TANH" fait référence à la fonction d'activation de type tangente hyperbolique. Des mesures objectives ont été calculées sur les stimuli synthétisés avec les différentes architectures. Les résultats obtenus sont représentés dans le tableau 5.2. Les résultats de l'évaluation objective montrent que l'utilisation des BLSTM permet d'obtenir la meilleure qualité de parole générée, ceci est dû au fait que cette architecture tient compte de l'aspect séquentiel de la parole (section 2.4.2).

TABLE 5.2 – Mesure objectives

Mesures objectives	DNN	LSTM	GRU	BLSTM
MCD (dB)	4,4	4,3	4,4	**4,3**
BAP (dB)	0,3	0,3	0,3	**0,3**
RMSE F0 (Hz)	11,1	11,2	11,0	**10,8**
VUV (%)	5,8	4,0	4,1	**4,0**
RMSE (ms)	27,5	28,0	29,0	**28,0**

Dans la suite des expériences, l'alignement par état a été utilisé. Le processus de synthèse a été configuré comme suit :

— Architecture : BLSTM

— Vocodeur : WORLD

— Algorithme d'optimisation : adam

5.4 Modélisation des unités de parole

Dans cette partie, l'impact des quatre approches de modélisation (C1V1, C1V2, C2V1 et C2V2) sur la qualité des signaux générés est évalué. Ceci permettra de déduire si les DNN tirent profit de la différentiation des unités modélisant les consonnes simples de celles géminées d'un part et la différentiation des voyelles courtes des longues d'autre part (Houidhek *et al.*, 2018a).

5.4.1 Évaluation objective

Pendant les expériences, la même répartition des ensembles des tests et d'apprentissage que celle décrite dans la section 5.3 a été utilisée : un ensemble de 1565 phrases a été utilisé pour l'apprentissage et 30 phrases ont été gardés pour le test. L'alignement par état et l'architecture BLSTM qui ont permis d'avoir les meilleurs mesures objectives ont été utilisés (cf. section 5.3.2).

Évaluation objective des durées de phonèmes

Les durées prédites des sons ont été jugées en calculant la moyenne sur l'ensemble des voyelles des rapports entre la durée moyenne des voyelles longues (VL) et la durée

moyenne des voyelles courtes équivalentes (VC), ainsi que la moyenne sur l'ensemble des consonnes des rapports entre les durées moyennes des consonnes géminées (CG) et celles des consonnes simples (CS). Seuls les phonèmes avec plus de 10 occurrences pour chaque classe (consonnes simples, consonnes géminées, voyelles courtes et voyelles longues) sont considérés. Les rapports calculés sont comparés à ceux obtenus pour la parole naturelle. Les résultats obtenus sont représentés dans le tableau 5.3.

TABLE 5.3 – Rapports des durées

Nombre d'occurences	**VL / VC 262 / 884**	**CG / SC 104 / 1315**
$C1V1$	1,7	2,1
$C1V2$	1,7	2,1
$C2V1$	1,8	2,2
$C2V2$	1,7	2,1
Original	2,0	2,1

Les valeurs montrent que pour les quatre approches de modélisation, les rapports entre les durées prédites des voyelles longues (VL) et des voyelles courtes (VC) et les rapports entre les durées prédites des consonnes géminées (CG) et les durées prédites des consonnes simples (CS) sont similaires à ceux calculés sur la parole naturelle.

L'erreur quadratique moyenne (RMSE) entre la durée naturelle et celle prédite a été calculée pour les différentes classes de phonèmes. Les valeurs de RMSE sont présentées dans la figure 5.1. Les résultats montrent que pour chaque classe de phonèmes, le modèle C2V2 (le modèle le plus détaillé) est caractérisé par une valeur de RMSE plus faible que celles obtenues pour les autres modèles (C1V1, C1V2 et C2V1).

Pour compléter, l'erreur quadratique normalisée a été calculée en considérant les valeurs de durée moyenne de chaque classe de phonèmes (NRMSE = RMSE / durée moyenne). Les résultats obtenus sont présentés sur la figure 5.2. Le NRMSE du modèle C2V2 présente une diminution significative par rapport au NRMSE des autres modèles qui, pour chaque classe de phonèmes, présentent des valeurs similaires de NRMSE.

Évaluation objective des paramètres acoustiques

Des mesures objectives ont été calculées afin d'évaluer la prédiction des paramètres acoustiques des signaux générés par les quatre modèles. Il s'agit du calcul de la distorsion des

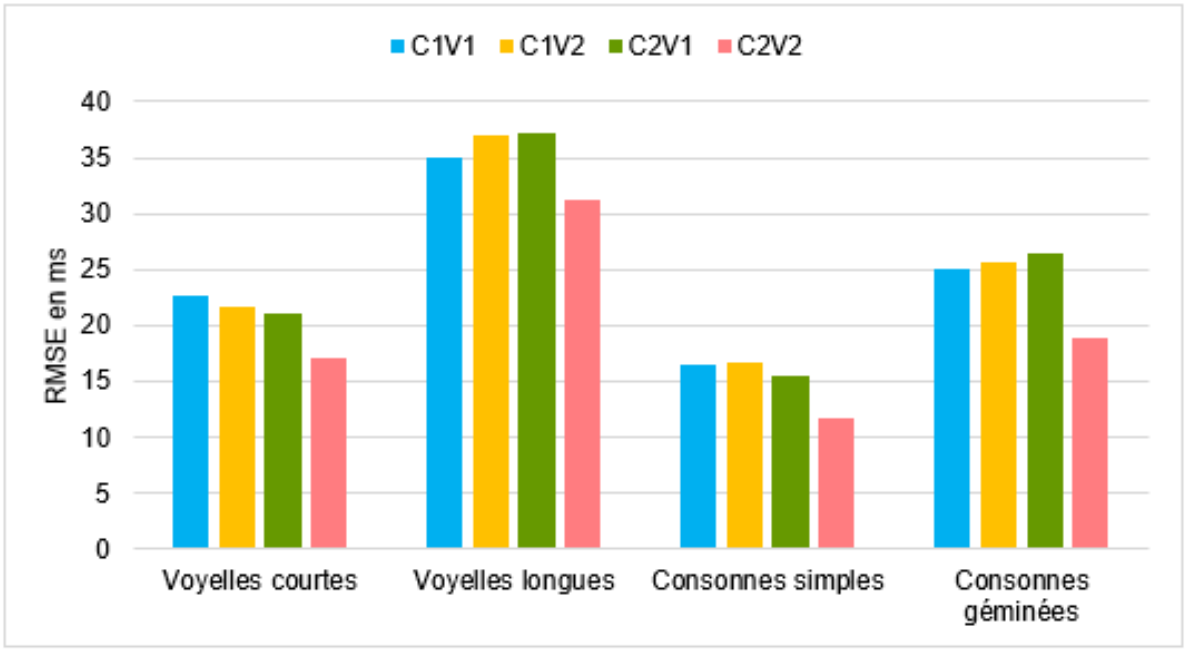


FIGURE 5.1 – RMSE entre durée naturelle et durée prédite

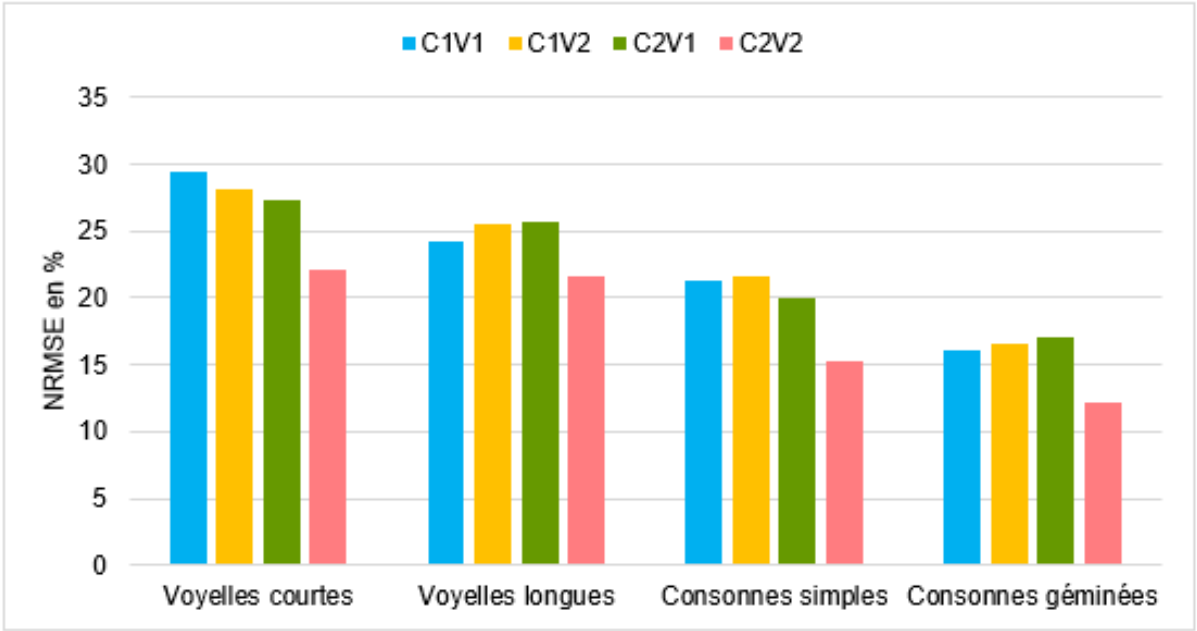


FIGURE 5.2 – NRMSE entre durée naturelle et durée prédite

coefficients spectraux (MCD (db)), de l'erreur de prédiction de la fréquence fondamentale (F0), de la bande d'apériodicité (BAP) et du voisement (VUV) entre les signaux générés et les signaux naturels correspondants. Les valeurs obtenues sont représentées dans le tableau 5.4

TABLE 5.4 – Évaluation de la prédiction des paramètres acoustiques

Approche	F0 (Hz)	BAP (dB)	VUV (%)	MCD (dB)
$C1V1$	14,78	0,28	3,60	4,02
$C1V2$	14,63	0,28	3,60	4,02
$C2V1$	13,68	0,27	3,62	3,99
$C2V2$	13,76	0,27	3,65	3,94

Les valeurs obtenues montrent que les quatre modèles (C2V2, C2V1, C1V2, C1V1) permettent d'obtenir des valeurs similaires de l'erreur de prédiction de la fréquence fondamentale et des paramètres spectraux. Ceci permet de conclure que les signaux synthétisés avec les quatre modèles, devraient présenter une intonation et un rythme similaires.

5.4.2 Évaluation subjective

Un test perceptif de préférence a été conduit afin de comparer les signaux générés avec les quatre modèles (ITU, 1996).

Données évaluées

Pour le test, les phrases suivantes ont été considérées :

— les signaux synthétisés par DNN en utilisant le modèle C1V1.

— les signaux synthétisés par DNN en utilisant le modèle C1V2.

— les signaux synthétisés par DNN en utilisant le modèle C2V1.

— les signaux synthétisés par DNN en utilisant le modèle C2V2.

Pendant l'évaluation, les quatre systèmes sont comparés deux à deux.

Protocole

Dix-huit locuteurs natifs arabes ont participé à cette évaluation. Chacun a évalué un ensemble de 20 paires de signaux vocaux ; chaque paire est constituée du même énoncé produit avec deux modèles différents. L'ordre de présentation des signaux de parole est choisi au hasard. Au cours de l'évaluation, les participants étaient appelés à indiquer le signal préféré selon la qualité globale de la parole produite, en répondant à la question "Comment jugez-vous la qualité du second signal par rapport au premier ?" par une note allant de 1, "beaucoup plus mauvaise" à 7, "Bien meilleure". Afin de faciliter l'analyse des résultats, les scores ont été regroupés en trois catégories : « premier préféré », « pas de préférence » et « deuxième préféré», comme suit (cf. Annexe F) :

$$\left.\begin{array}{ll} 1 & \textbf{Beaucoup plus mauvaise} \\ 2 & \textbf{Plus mauvaise} \end{array}\right\} = \textbf{Premier préféré}$$

$$\left.\begin{array}{ll} 3 & \textbf{Un peu plus mauvaise} \\ 4 & \textbf{A peu près la même} \\ 5 & \textbf{Un peu meilleure} \end{array}\right\} = \textbf{Pas de préférence}$$

$$\left.\begin{array}{ll} 6 & \textbf{Meilleure} \\ 7 & \textbf{Bien meilleure} \end{array}\right\} = \textbf{Deuxième préféré}$$

Résultats

Les scores obtenus sont représentés dans la figure 5.3. Les résultats des tests perceptifs montrent que les auditeurs n'ont pas perçu de différence significative entre les signaux générés par les quatre modèles. Ceci malgré le fait que les évaluations objectives ont montré que le modèle C2V2 permet d'avoir la meilleure prédiction de la durée, de la fréquence fondamentale et des paramètres spectraux.

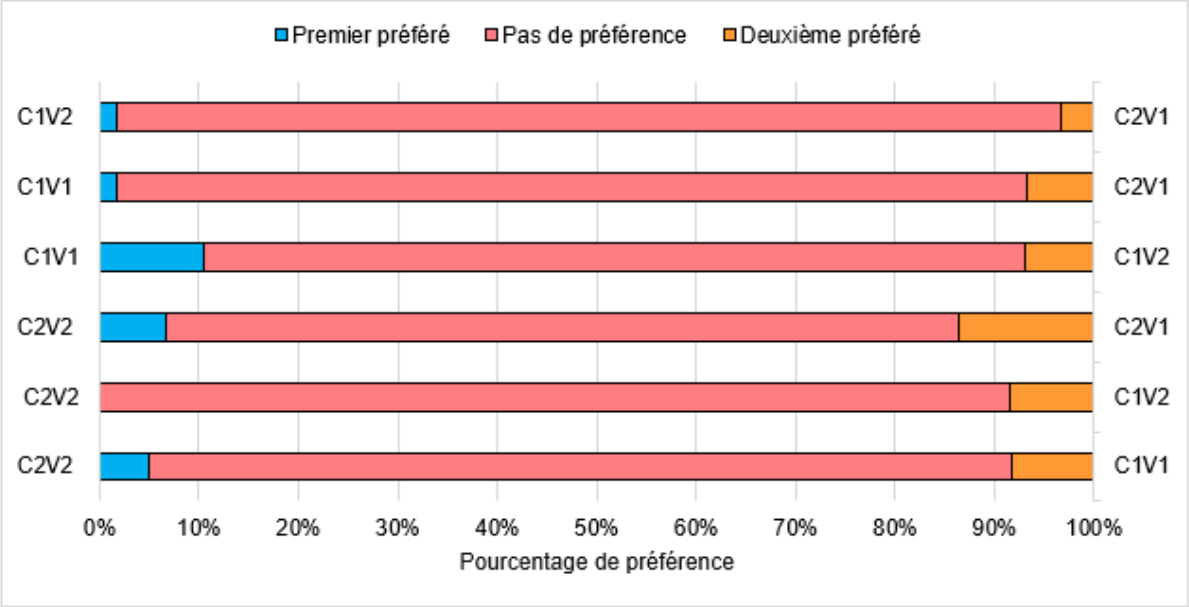


FIGURE 5.3 – Comparaison des quatre modèles

5.5 Comparaison de la synthèse par HMM et par DNN

Des tests perceptifs ont été conduits afin de comparer les performances de la synthèse de parole arabe par HMM et par DNN. La qualité globale et l'aspect naturel des signaux synthétisés avec les HMM et les DNN ont été évalués. Ensuite, les taux de dégradation perçus des signaux de parole générés par HTS et par MERLIN en comparaison de la parole naturelle ont été mesurés, ceci en comparant les phrases produites aux phrases naturelles.

5.5.1 Évaluation objective de la durée

Dans cette partie, les mesures objectives faites à propos de la prédiction des durées des segments de parole en utilisant les différentes approches de synthèse (HMM, DNN) sont comparées.

Rapports des durées : La moyenne sur l'ensemble des voyelles des rapports entre la durée moyenne des voyelles longues (VL) et la durée moyenne des voyelles courtes équivalentes (VC), ainsi que la moyenne sur l'ensemble des consonnes des rapports entre les durées moyennes des consonnes géminées (CG) et celles des consonnes

simples (CS). Seuls les phonèmes avec plus de 10 occurrences pour chaque classe (consonnes simples, consonnes géminées, voyelles courtes et voyelles longues) sont considérés. Les rapports calculés sont comparés à ceux obtenus pour la parole naturelle. Les résultats obtenus sont représentés dans le tableau 5.5.

TABLE 5.5 – Rapport des durées

	HMM		DNN	
Nombre d'occurences	VL / VC 262 / 884	CG / SC 104 / 1315	VL / VC 262 / 884	CG / SC 104 / 1315
$C1V1$	$1,7$	$2,1$	1,7	2,1
$C1V2$	$1,7$	$2,1$	1,7	2,1
$C2V1$	$1,7$	$2,1$	1,8	2,2
$C2V2$	$1,8$	$2,2$	1,7	2,1
Originale	$2,0$	$2,1$	2,0	2,1

Les rapports obtenus montrent une similarité des valeurs dans les deux approches de synthèse : par HMM et par DNN.

RMSE entre durée naturelle et durée prédite : L'erreur quadratique moyenne (RMSE) entre la durée naturelle et celle prédite a été calculée pour les différentes classes de phonèmes dans le cas de synthèse par HMM et par DNN. Les valeurs de RMSE sont présentées dans le tableau 5.6. Les valeurs montrent que pour chaque approche de modélisation, les DNN permettent de mieux prédire les durées des phonèmes de chaque classe (consonnes simples, consonnes géminées, voyelles courtes, voyelles longues).

TABLE 5.6 – RMSE entre durée naturelle et durée prédite

	C1V1		C1V2		C2V1		C2V2	
	DNN	HMM	DNN	HMM	DNN	HMM	DNN	HMM
Voyelles courtes	23	27	22	28	22	28	15	27
Voyelles longues	35	49	37	45	37	50	30	44
Consonnes simples	16	23	16	23	15	23	15	24
Consonnes géminées	25	32	26	34	26	36	20	37

NRMSE entre durée naturelle et durée prédite : L'erreur quadratique normalisée a été calculée en considérant les valeurs de durée moyenne de chaque classe de

phonèmes (NRMSE = RMSE / durée moyenne) dans le cas de synthèse de parole par HMM et par DNN. Les résultats obtenus sont présentés dans le tableau 5.7. Ces résultats confirment que les DNN présentent une meilleure performance au niveau de la prédiction des durées des segments de parole.

TABLE 5.7 – NRMSE entre durée naturelle et durée prédite

	C1V1		C1V2		C2V1		C2V2	
	DNN	HMM	DNN	HMM	DNN	HMM	DNN	HMM
Voyelles courtes	29	33	28	34	27	34	22	34
Voyelles longues	24	33	26	30	26	34	22	30
Consonnes simples	21	32	22	32	20	32	15	33
Consonnes géminées	16	21	16	22	17	23	12	24

5.5.2 Évaluation de la qualité globale et de l'aspect naturel

La qualité globale fait référence à la qualité du signal acoustique. L'aspect naturel est évalué en se référant à l'intonation et au rythme des signaux vocaux synthétisés.

Données évaluées

Pendant les expériences, les signaux suivants ont été évalués :

— phrases synthétisées par HMM en utilisant le toolkit HTS.

— phrases synthétisées par DNN en utilisant le toolkit MERLIN.

Dans les deux cas, seul le modèle C2V2 est considéré. Pour chaque ensemble, 30 phrases ont été générées. Il s'agit des mêmes signaux que ceux utilisés dans les chapitres précédents.

Protocole

Des tests MOS (ITU, 1996) ont été conduits. La question posée aux participants est "Comment évaluez-vous la qualité globale et l'aspect naturel de ce que vous venez d'entendre par rapport à une parole naturelle (prononcée par un être humain) ?» L'évaluation se fait en donnant un score après l'écoute du signal selon l'échelle suivante :

1. Très loin de la parole naturelle.
2. Loin de la parole naturelle

3. Un peu loin de la parole naturelle.
4. Proche de la parole naturelle.
5. Très proche de la parole naturelle.

Quinze auditeurs ont participé à l'évaluation, et chacun a évalué un ensemble de 20 phrases : 10 phrases générées par chaque approche (DNN ou HMM) (cf. Annexe D).

Résultats

Les scores MOS obtenus sont représentés dans la figure 5.4 associés à leurs intervalles de confiance à 95%. Les résultats montrent que les phrases générées avec les DNN ont obtenu les scores les plus élevés comparés à ceux obtenus par les phrases générées par les HMM.

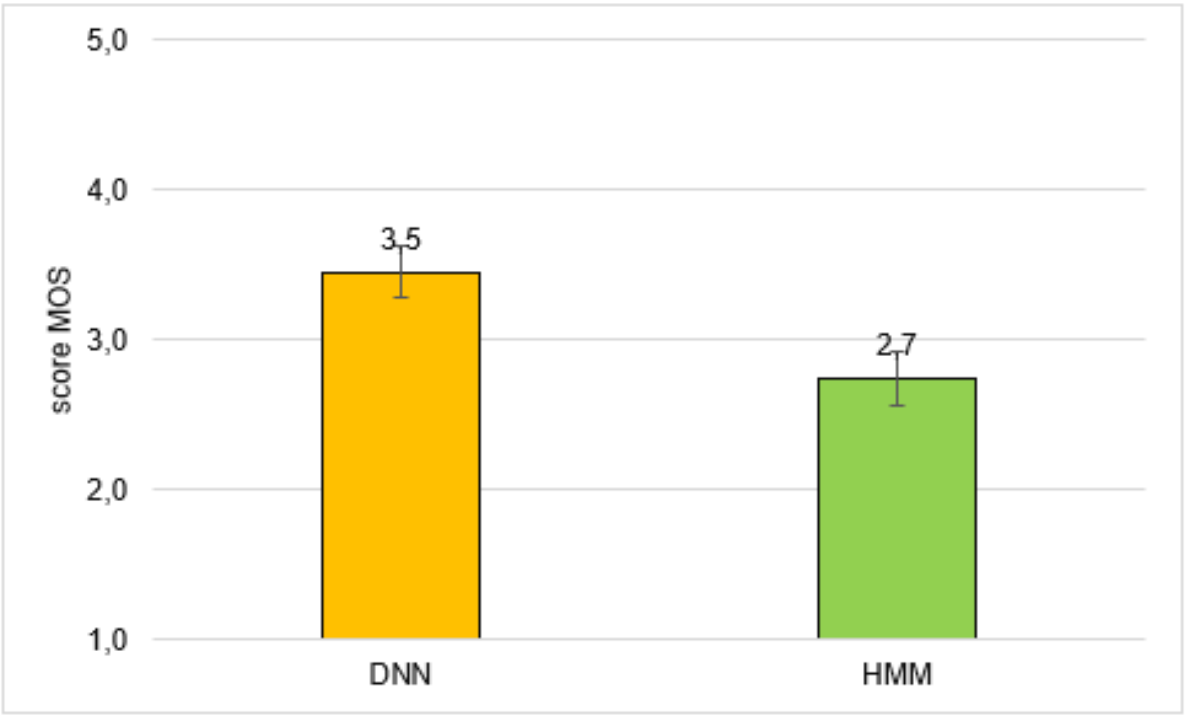


FIGURE 5.4 – Évaluation de la qualité globale

5.5.3 Résultats du test DMOS

Ce test a pour but la mesure du degré de la dégradation engendré par le système de synthèse de parole (ITU, 1996). Il s'agit de mesurer la dégradation perçue pour les signaux

synthétisés en comparaison aux signaux naturels correspondants.

Données évaluées

Pour ce test, les signaux suivants sont utilisés :

— signaux synthétisés par HMM

— signaux synthétisés par DNN

— signaux reconstruits avec la procédure analyse/synthèse en utilisant le vocodeur WORLD.

Ces signaux ont été comparés aux signaux naturels correspondants.

Protocole

Neuf auditeurs ont participé à ce test. Les participants ont été amenés à juger le taux de la dégradation en répondant à la question : Comment évaluez-vous la dégradation du deuxième signal par rapport au premier ? En utilisant une échelle à cinq points de dégradation (cf. Annexe E) :

1. Dégradation très gênante
2. Dégradation gênante
3. Dégradation un peu gênante
4. Dégradation audible mais pas gênante
5. Dégradation inaudible

Chaque participant a évalué un ensemble de 30 paires de phrases (10 provenant de la synthèse par HMM, 10 de la synthèse par DNN et 10 reconstruits par analyse/synthèse). L'autre élément de la paire est le signal naturel correspondant.

Résultats

Les scores DMOS obtenus ont été représentés dans la figure 5.5 avec les intervalles de confiance associés à 95%.

Plus le score est élevé, plus la dégradation est faible. Les scores obtenus montrent que les signaux générés en utilisant les DNN présentent le même degré de dégradation que les signaux reconstruits avec le vocodeur WORLD en comparaison aux signaux naturels.

Les signaux générés par les HMM présentent un taux de dégradation relativement plus important.

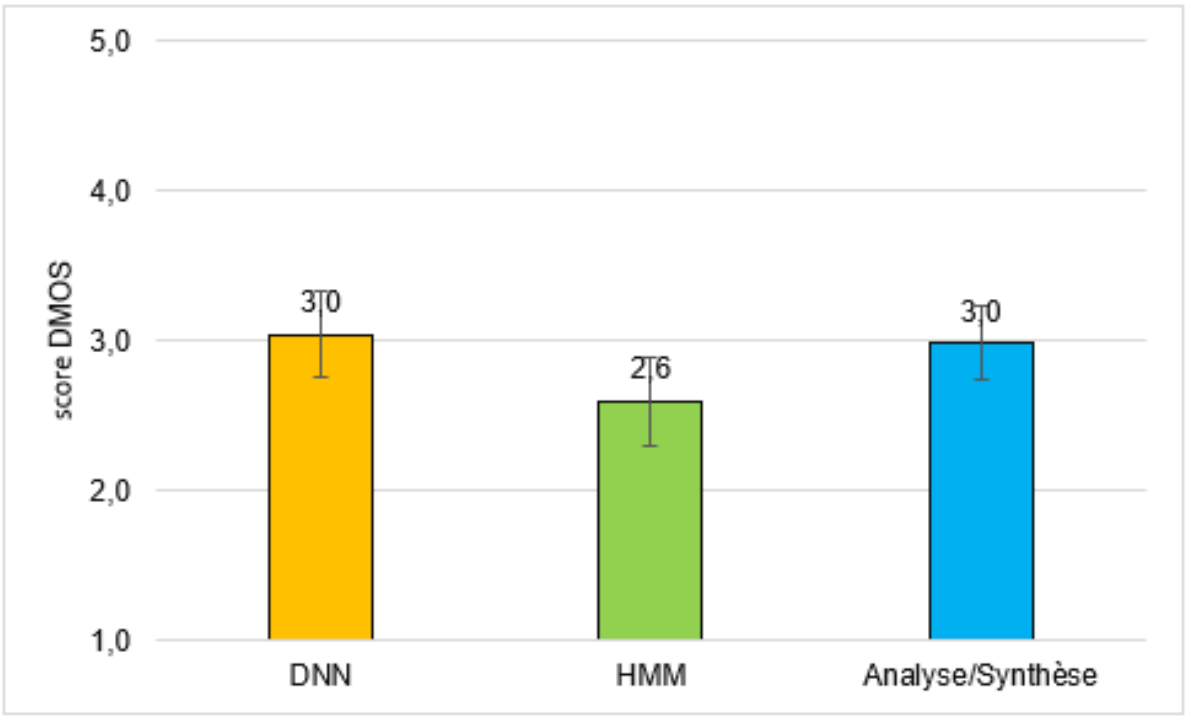


FIGURE 5.5 – Évaluation de la dégradation

5.5.4 Comparaison HMM vs. DNN

Il s'agit d'un test de préférence (ITU, 1996) permettant de comparer les performances des HMM à celles des DNN pour la synthèse de la parole arabe.

Données évaluées

Pour ce test, les signaux suivants sont utilisés :

— signaux synthétisés par HMM

— signaux synthétisés par DNN

— signaux reconstruits avec la procédure analyse/synthèse en utilisant le vocodeur WORLD.

Protocole

La comparaison des signaux se fait selon la qualité globale. Dix-huit auditeurs ont participé à ce test ; chacun a évalué un ensemble de 30 paires de phrases. Chaque paire est formée par deux phrases synthétisées par deux systèmes différents ou une phrase produite par un système de synthèse et une reconstruite par analyse / synthèse. L'ordre de présentation des signaux est complètement aléatoire. Les participants ont été appelés à indiquer leur préférence en répondant à la question suivante : "Comment jugez-vous la qualité du second signal par rapport au premier ?" La réponse est un score allant de 1, "beaucoup plus mauvaise" à 7, "Bien meilleure".

Afin de faciliter l'analyse des résultats, les scores ont été regroupés en trois catégories : « premier préféré », « pas de préférence » et « deuxième préféré », comme suit :

$$\left.\begin{array}{ll} 1 & \textbf{Beaucoup plus mauvaise} \\ 2 & \textbf{Plus mauvaise} \end{array}\right\} = \textbf{Premier préféré}$$

$$\left.\begin{array}{ll} 3 & \textbf{Un peu plus mauvaise} \\ 4 & \textbf{A peu près la même} \\ 5 & \textbf{Un peu meilleure} \end{array}\right\} = \textbf{Pas de préférence}$$

$$\left.\begin{array}{ll} 6 & \textbf{Meilleure} \\ 7 & \textbf{Bien meilleure} \end{array}\right\} = \textbf{Deuxième préféré}$$

Résultats

Les scores obtenus après regroupement sont représentés dans la figure 5.6 :

Les résultats obtenus montrent que les signaux générés par DNN et ceux reconstruits en utilisant le vocodeur WORLD sont préférés par rapport aux signaux synthétisés par HMM. Ceci confirme le fait que l'utilisation des DNN pour assurer le passage des descripteurs contextuels aux paramètres acoustiques est plus efficace que l'utilisation des HMM.

5.6 Conclusion

Ce chapitre a étudié l'utilisation des DNN pour la synthèse paramétrique de la parole arabe. Une première partie a été consacrée à l'évaluation de l'impact des quatre approches

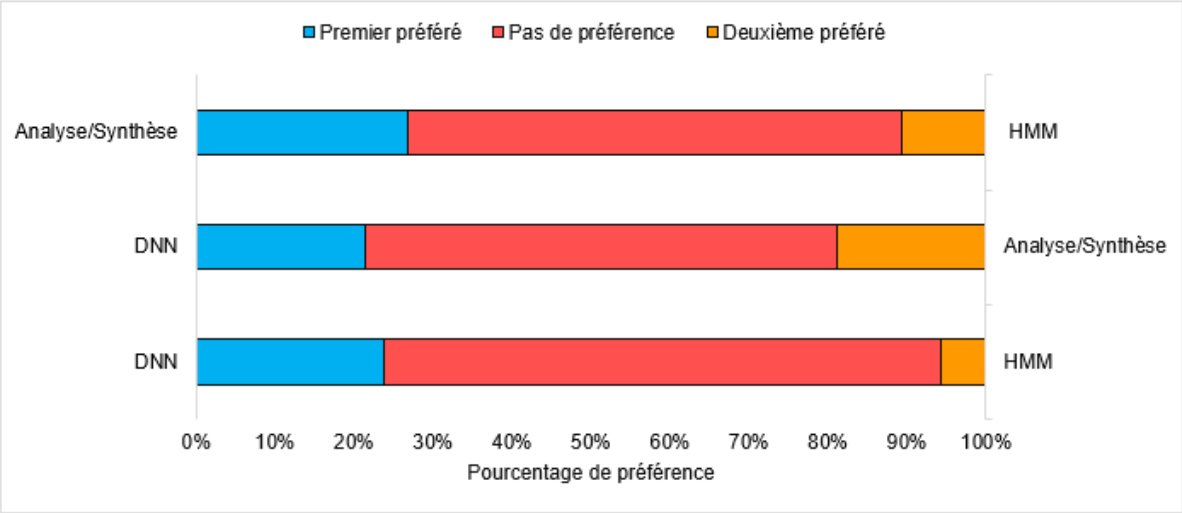


FIGURE 5.6 – Résultats des tests de comparaison

de modélisation sur la qualité globale des signaux. Les tests perceptifs ont montré que les auditeurs n'ont pas perçu de différence significative entre les signaux de parole à l'écoute, cependant les évaluations objectives ont montré que le modèle C2V2 permet de mieux prédire la durée et la fréquence fondamentale par rapport aux autres modèles.

La deuxième partie a été dédiée à l'évaluation des performances des approches par HMM et par DNN pour modéliser le passage des descripteurs contextuels aux paramètres acoustiques. Les résultats de l'évaluation subjective ont montré que les DNN sont plus efficaces pour la prédiction des paramètres acoustiques ; les DNN ont amélioré la précision du modèle acoustique.

Conclusion générale

Le but des travaux présentés dans cette thèse était l'application de l'approche de synthèse paramétrique à la parole arabe en utilisant les HMM et les DNN.

L'adaptation de l'approche paramétrique de synthèse nécessite une qualification de chaque segment de parole par un ensemble des descripteurs contextuels dont une partie dépend de la langue. De ce fait, nous avons tout d'abord étudié les caractéristiques de la langue arabe afin de définir l'ensemble des descripteurs contextuels qui caractérisent les différents contextes dans lesquels une unité de parole peut exister. On a mis l'accent sur deux phénomènes phonologiques : la gémination et les voyelles longues. On a ajouté deux descripteurs se référant aux informations de gémination et à la longueur des voyelles à l'ensemble standard des descripteurs. Un ensemble des règles a été utilisé afin de prédire la position de l'accentuation dans le mot.

Une partie du travail a porté sur le choix de la modélisation des consonnes géminées et des voyelles longues par rapport aux consonnes simples et voyelles courtes. Quatre approches de modélisation ont été proposées

— Différentiation des consonnes géminées (respectivement voyelles longues) des consonnes simples (respectivement voyelles courtes)

— Fusion des consonnes géminées (respectivement voyelles longues) avec les consonnes simples (respectivement voyelles courtes)

— Différentiation des consonnes géminées et des consonnes simples et fusion des voyelles courtes avec les voyelles longues.

— Fusion des consonnes simples avec les consonnes géminées et différentiation des voyelles courtes et voyelles longues.

Dans la première partie, l'approche de synthèse paramétrique statistique (par HMM) a été utilisée pour convertir le texte arabe en des signaux de parole. Le système HTS a été utilisé pour la mise en œuvre. Les quatre approches de modélisation ont été développées et utilisées pour générer un ensemble de signaux de parole qui ont été évalués par la suite par des mesures subjectives et objectives. Des mesures objectives ont été calculées afin d'évaluer les durées prédites par les quatre approches de modélisation ceci en calculant la moyenne sur l'ensemble des voyelles des rapports entre la durée moyenne des voyelles longues et la durée moyenne des voyelles courtes équivalentes, ainsi que la moyenne, sur l'ensemble des consonnes, des rapports entre les durées moyennes des consonnes géminées et celles des consonnes simples. En comparant les rapports obtenus à ceux calculés pour la parole naturelle, il a été observé qu'il y a une similarité entre les rapports des durées pour les quatre approches de modélisation et ceux de la parole naturelle. L'évaluation subjective avait pour but le jugement de l'effet des quatre approches de modélisation sur la qualité globale et l'aspect naturel des signaux générés. Un protocole d'évaluation a été défini et trois types de tests ont été conduits : test MOS de qualité globale, test DMOS pour l'évaluation du degré de dégradation causé par le système de synthèse et un dernier test de comparaison. Les résultats obtenus ont montré qu'il n'y avait pas de différence importante entre les différentes approches de modélisation. Par conséquent, l'identification des consonnes géminées et des voyelles longues en tant que phonèmes à part entière (modélisés par des unités spécifiques) n'est pas nécessaire lors de l'utilisation de HTS pour la synthèse de la parole arabe, tant que ces informations existent dans l'ensemble des caractéristiques contextuelles.

Dans la deuxième partie, l'approche de synthèse paramétrique par DNN a été étudiée et le toolkit MERLIN a été utilisé pour générer les signaux de parole arabes. Le même ensemble des descripteurs contextuels que ceux utilisés dans l'expérience précédente a été adopté. De même, les quatre approches de modélisation ont été employées pour la production des signaux de parole. Dans cette partie, deux types d'évaluations ont été conduites ; tout d'abord une comparaison des quatre approches de modélisation, ensuite une comparaison des performances des HMM et des DNN quand ils sont utilisés pour la synthèse de la parole arabe.

Pendant l'évaluation objective des durées prédites par les quatre approches de modélisation, les mêmes rapports que ceux décrits précédemment ont été calculés. Les résultats obtenus ont montré que pour les quatre approches de modélisation, les rapports entre

les durées prédites des voyelles longues et des voyelles courtes et ceux calculés entre les durées prédites des consonnes géminées et les durées prédites des consonnes simples sont similaires à ceux calculés sur la parole naturelle. Nous avons également utilisé un autre critère, l'erreur quadratique moyenne et l'erreur normalisée, l'approche de modélisation la plus détaillée a été distinguée des autres approches par une erreur réduite, ce qui mène à dire que cette approche permet une meilleure prédiction des durées par rapport aux autres approches. Lors des tests perceptifs, les auditeurs n'étaient cependant pas capables de percevoir cette différence entre les approches, les quatre approches ont été jugées similaires en terme de qualité globale et pour le test de préférence.

Finalement, la dernière partie a été consacrée à la comparaison des performances des HMM et des DNN pour modéliser le passage des descripteurs contextuels aux paramètres acoustiques. Un ensemble des tests perceptifs a été conduit afin de comparer les effets de l'utilisation des DNN et des HMM sur la qualité globale, l'aspect naturel et la dégradation des signaux générés. Les résultats de l'évaluation subjective ont montré que les DNN sont plus efficaces pour la prédiction des paramètres acoustiques ; les DNN ont amélioré la précision du modèle acoustique.

Annexe A

Les descripteurs contextuels standards

Chaque phonème est décrit par les informations suivantes :

```
    p1^p2-p3+p4=p5@p6_p7
/A:a1_a2_a3/B:b1-b2-b3@b4-b5&b6-b7#b8-b9$b10-b11!b12-b13;b14-b15|b16/C:c1+c2+c3
/D:d1_d2/E:e1+e2@e3+e4&e5+e6#e7+e8/F:f1_f2
/G:g1_g2/H:h1=h2@h3=h4|h5/I:i1_i2
/J:j1+j2-j3
```

Index	Label	Description du descripteur
1	p1	phonème précédent-précédent
2	p2	phonème précédent
3	p3	phonème courant
4	p4	phonème suivant
5	p5	phonème suivant-suivant
6	p6	position du phonème courant dans la syllabe courante (à partir du début)
7	p7	position du phonème courant dans la syllabe courante (à partir de la fin)
8	a1	est ce que la syllabe précédente porte un accent lexical ?
9	a2	est ce que la syllabe précédente est accentuée ?
10	a3	nombre de phonèmes dans la syllabe précédente
11	b1	est ce que la syllabe courante porte un accent lexical ?
12	b2	est ce que la syllabe courante est accentuée ?
13	b3	nombre de phonèmes dans la syllabe courante
14	b4	position de la syllabe courante dans le mot courant (à partir du début)

15	b5	position de la syllabe courante dans le mot courant (à partir de la fin)
16	b6	position de la syllabe courante dans la phrase courante (à partir du début)
17	b7	position de la syllabe courante dans la phrase courante (à partir de la fin)
18	b8	nombre de syllabes stressée avant la syllabe courante dans la phrase courante
19	b9	nombre de syllabes stressée après la syllabe courante dans la phrase courante
20	b10	nombre de syllabes accentuée avant la syllabe courante dans la phrase courante.
21	b11	nombre de syllabes accentuées après la syllabe courante dans la phrase courante.
22	b12	nombre de syllabes à partir de la dernière syllabe accentuée jusque-là syllabe courante.
23	b13	nombre de syllabes à partir de la syllabe courante jusque-là prochaine syllabe stressée.
24	b14	nombre de syllabes à partir de la dernière syllabe accentuée jusqu'à la syllabe courante.
25	b15	nombre de syllabes à partir de la syllabe courante jusque-là prochaine syllabe accentuée.
26	b16	label de la voyelle de la syllabe courante.
27	c1	est ce que la syllabe suivante accentuée ?
28	c2	est ce que la syllabe suivante stressée ?
29	c3	nombre de phonèmes dans la syllabe suivante.
30	d1	classe grammaticale du mot précédent.
31	d2	nombre de syllabes dans le mot précédent.
32	e1	classe grammaticale du mot courant.
33	e2	nombre de syllabes dans le mot courant.
34	e3	position du mot courant dans la phrase courante (à partir du début).
35	e4	position du mot courant dans la phrase courante (à partir de la fin).
36	e5	nombre de mots signifiants avant le mot courant dans la phrase courante.
37	e6	nombre de mots signifiants après le mot courant dans la phrase courante.
38	e7	nombre de mots à partir du dernier mot portant un accent jusqu'au mot courant.
39	e8	nombre de mots à partir du mot courant jusqu'au prochain mot portant un accent.
40	f1	classe grammaticale du mot suivant.
41	f2	nombre de syllabes dans le mot suivant.
42	g1	nombre de syllabes dans la phrase précédente.
43	g2	nombre de mots dans la phrase précédente
44	h1	nombre de syllabes dans la phrase courante
45	h2	nombre de mots dans la phrase courante

46	h3	position de la phrase dans l'énoncé (à partir du début)
47	h4	position de la phrase dans l'énoncé (à partir de la fin)
48	h5	Tag ToBi de fin de phrase
49	i1	nombre de syllabes dans la phrase suivante
50	i2	nombre de mots dans la phrase suivante
51	j1	nombre de syllabes dans l'énoncé
52	j2	nombre de mots dans l'énoncé
53	j3	nombre de phrases dans l'énoncé

Annexe B

Ensemble des descripteurs utilisés

Index	Label	Description du descripteur
1	p1	phonème précédent-précédent
2	p2	phonème précédent
3	p3	phonème courant
4	p4	phonème suivant
5	p5	phonème suivant-suivant
6	p6	position du phonème courant dans la syllabe courante (à partir du début)
7	p7	position du phonème courant dans la syllabe courante (à partir de la fin)
9	a2	est ce que la syllabe précédente est accentuée ?
10	a3	nombre de phonèmes dans la syllabe précédente
11	b1	est ce que la syllabe courante porte un accent lexical ?
13	b3	nombre de phonèmes dans la syllabe courante
14	b4	position de la syllabe courante dans le mot courant (à partir du début)
15	b5	position de la syllabe courante dans le mot courant (à partir de la fin)
16	b6	position de la syllabe courante dans la phrase courante (à partir du début)
17	b7	position de la syllabe courante dans la phrase courante (à partir de la fin)
18	b8	nombre de syllabes stressée avant la syllabe courante dans la phrase courante
19	b9	nombre de syllabes stressée après la syllabe courante dans la phrase courante
20	b10	nombre de syllabes accentuée avant la syllabe courante dans la phrase courante.
23	b13	nombre de syllabes à partir de la syllabe courante jusque-là prochaine syllabe stressée.

26	b16	label de la voyelle de la syllabe courante.
28	c2	est ce que la syllabe suivante stressée ?
29	c3	nombre de phonèmes dans la syllabe suivante.
30	d1	classe grammaticale du mot précédent.
31	d2	nombre de syllabes dans le mot précédent.
32	e1	classe grammaticale du mot courant.
33	e2	nombre de syllabes dans le mot courant.
34	e3	position du mot courant dans la phrase courante (à partir du début).
35	e4	position du mot courant dans la phrase courante (à partir de la fin).
36	e5	nombre de mots grammaticaux avant le mot courant dans la phrase courante.
37	e6	nombre de mots grammaticaux après le mot courant dans la phrase courante.
38	e7	nombre de mots à partir du dernier mot portant un accent jusqu'au mot courant.
39	e8	nombre de mots à partir du mot courant jusqu'au prochain mot portant un accent.
40	f1	classe grammaticale du mot suivant.
41	f2	nombre de syllabes dans le mot suivant.
42	g1	nombre de syllabes dans la phrase précédente.
43	g2	nombre de mots dans la phrase précédente
44	h1	nombre de syllabes dans la phrase courante
45	h2	nombre de mots dans la phrase courante
46	h3	position de la phrase dans l'énoncé (à partir du début)
47	h4	position de la phrase dans l'énoncé (à partir de la fin)
49	i1	nombre de syllabes dans la phrase suivante
50	i2	nombre de mots dans la phrase suivante
51	j1	nombre de syllabes dans l'énoncé
52	j2	nombre de mots dans l'énoncé
53	j3	nombre de phrases dans l'énoncé
54	Z	Le phonème courant est-il une consonne simple ou géminée ou n'est pas une consonne
55	V	Le phonème courant est-il une voyelle courte, longue ou n'est pas une voyelle.

Annexe C

Consignes pour les évaluation subjectives

Une fois que l'auditeur accepte de faire le test d'écoute, il reçoit le lien contenant l'expérience. L'expérience commence par la présentation de consignes expliquant les objectifs et le déroulement du test.

Expérience de perception

Merci pour votre participation à cette étude...

L'objectif de cette expérience est l'étude de ...
Nous allons vous présenter des sons/vidéos de plusieurs locuteurs. Ces personnes vont prononcer soit des mots, soit des phrases complètes.

Afin que l'expérience se déroule dans les meilleures conditions, nous vous demanderons de **ne jamais utiliser le bouton précédent** de votre navigateur.

Le déroulement de l'expérience se fait comme suit :

1. Réglage du niveau sonore de votre ordinateur

2. Saisie d'informations (nom, prénom, âge, etc...)

   (ces informations resteront confidentielles et ne seront jamais divulguées.)

3. Lancement de l'évaluation.

   En règle générale, pour chaque son/vidéo:

   - **Ecoutez** et/ou **regardez** attentivement le son ou la vidéo.
   - **Cochez** ou **donnez** les réponses aux questions posées.
   - **Validez** votre/vos réponse(s). Le son ou la vidéo suivante seront présentés.

Il est primordial d'accorder la même attention à toutes les sons/vidéos afin que les résultats de l'étude soient fiables.

Etape 1: les réglages

L'étape suivante consiste en un réglage du matériel utilisé, tel que réglage du niveau sonore qui doit être fait une fois seulement avant de commencer le test.

Etape de réglage du niveau sonore

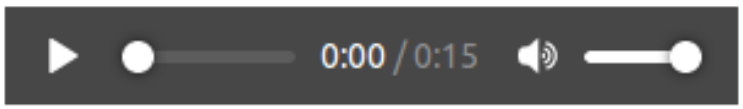
Voici la procédure à suivre:

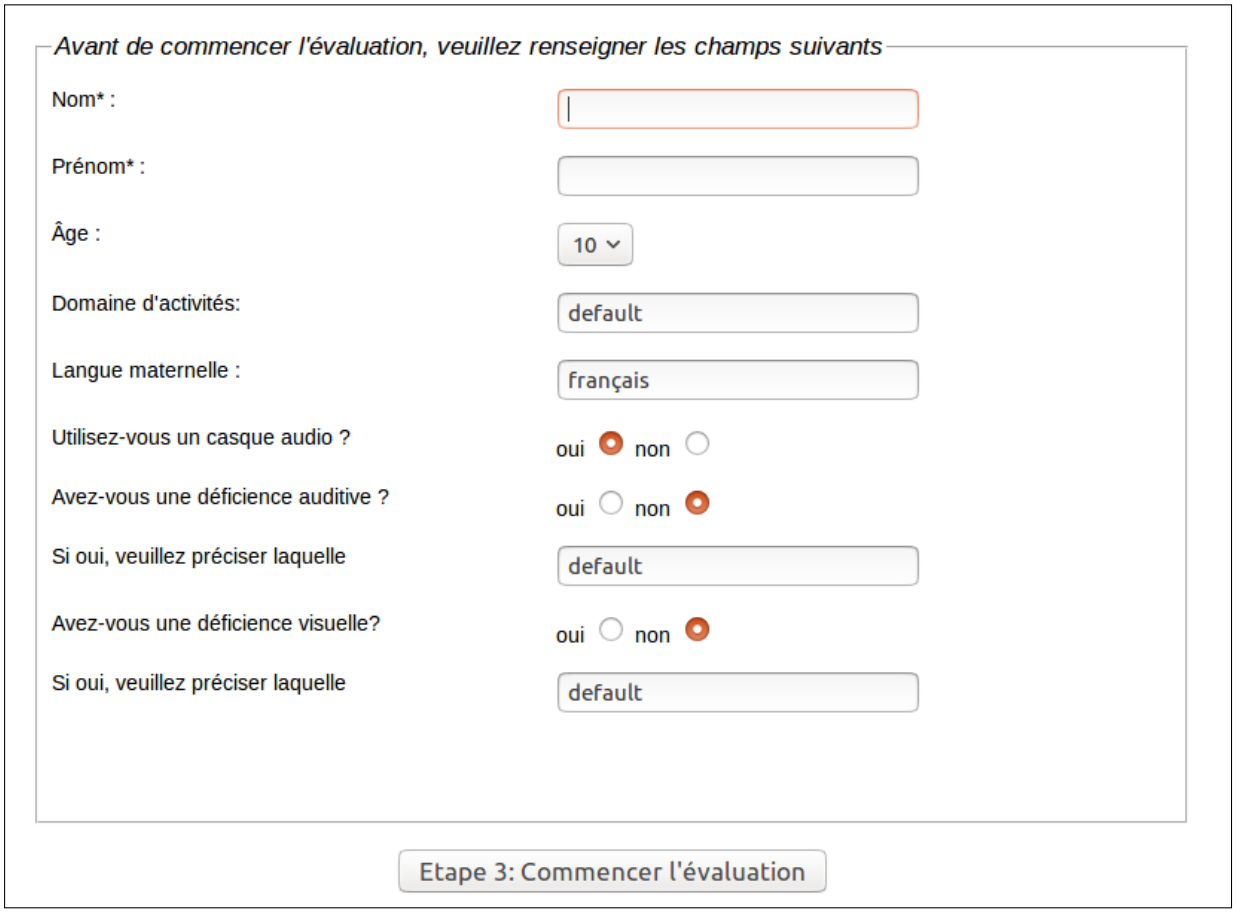
- Branchez **votre casque audio** si vous désirez l'utiliser pour l'évaluation.
  Nous vous recommandons son utilisation.
- Jouez la vidéo/le son ci-contre autant de fois que nécessaire.
- Réglez le volume de votre ordinateur de façon à ce que l'écoute soit la plus confortable possible.

Vous ne pouvez pas modifier le volume sur le lecteur vidéo, vous devez régler le volume de votre PC

Attention, il vous est demandé de ne pas modifier le volume une fois le test démarré
Si vous utilisez un casque il est important de le garder pour toute la durée du test.

Etape 2: Saisie d'information




Avant de commencer l'évaluation, veuillez renseigner les champs suivants
Nom* :
Prénom* :
Âge :
10
Domaine d'activités:
default
Langue maternelle :
français
Utilisez-vous un casque audio ?
oui
non
Avez-vous une déficience auditive ?
oui
non
Si oui, veuillez préciser laquelle
default
Avez-vous une déficience visuelle?
oui
non
Si oui, veuillez préciser laquelle
default
Etape 3: Commencer l'évaluation

Annexe D

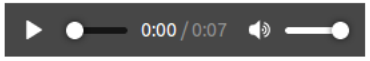
Test MOS (Mean Opinion Score)

Avant de commencer l'écoute des signaux, une brève introduction prend lieu, elle permet d'expliquer le but du test aux participants et comment juger les signaux et en se basant sur quel critère :

Dans cette expérience, vous allez juger l'aspect naturel d'un stimuli (à quel point il est naturel en termes d'intonation et de rythme, ceci revient à dire si le son, les longueurs des voyelles et des consonnes vous paraissent naturels) ainsi que sa qualité globale (en terme de qualité générale). Ainsi, après l'écoute du stimuli, veuillez l'évaluer en donnant un score de 1 à 5 selon l'échelle présentée.

Démarrer

Progression des evaluations: 1/1
Progression de cette étape : 1/33



En terme d'impression générale, comment jugez vous la qualité globale et l'aspect naturel de ce que vous venez d'écouter?

1	2	3	4	5
Mauvais	Médiocre	Passable	Bon	Excellent

Suivant

Annexe E

Test DMOS (Differential Mean Opinion Score)

Avant de commencer l'écoute des signaux générés et des signaux de référence, une brève introduction permet d'expliquer le but du test aux participants et comment juger les signaux et en se basant sur quel critère :

Dans ce test, vous allez évaluer le taux de la dégradation causée par le système de synthèse de la parole. Pour ce faire, vous allez tout d'abord faire l'écoute du stimuli de référence (parole naturelle), suivi de celui produit en utilisant un des systèmes de synthèse de la parole. Ensuite vous êtes appelés à juger la dégradation du deuxième stimuli par rapport au premier en attribuant un score de 1 à 5 selon l'échelle présentée.

Démarrer

Progression des evaluations: 1/1
Progression de cette étape : 1/32

A

B

Comment jugez-vous la dégradation du 2ème stimuli (B) par rapport au premier (A)?

1	2	3	4	5
Dégradation très gênante	Dégradation gênante	Dégradation un peu gênante	Dégradation audible mais pas gênante	Dégradation inaudible

Suivant

Annexe F

Test de préférence

Avant de comparer les signaux générés avec différentes approches, la démarche du test est expliquée aux auditeurs :

Dans cette expérience, vous allez faire l'écoute de deux stimuli dont chacun est synthétisé par une approche différente. Pour indiquer votre préférence en termes de qualité globale et intelligibilité (compréhensibilité) veuillez donner un score de 1 à 7 selon l'échelle présentée.

Démarrer

Progression des evaluations: 1/1
Progression de cette étape : 1/32

0:00 / 0:15
A

0:00 / 0:15
B

Comment jugez vous la qualité du deuxième stimuli (B) par rapport au premier (A) ?

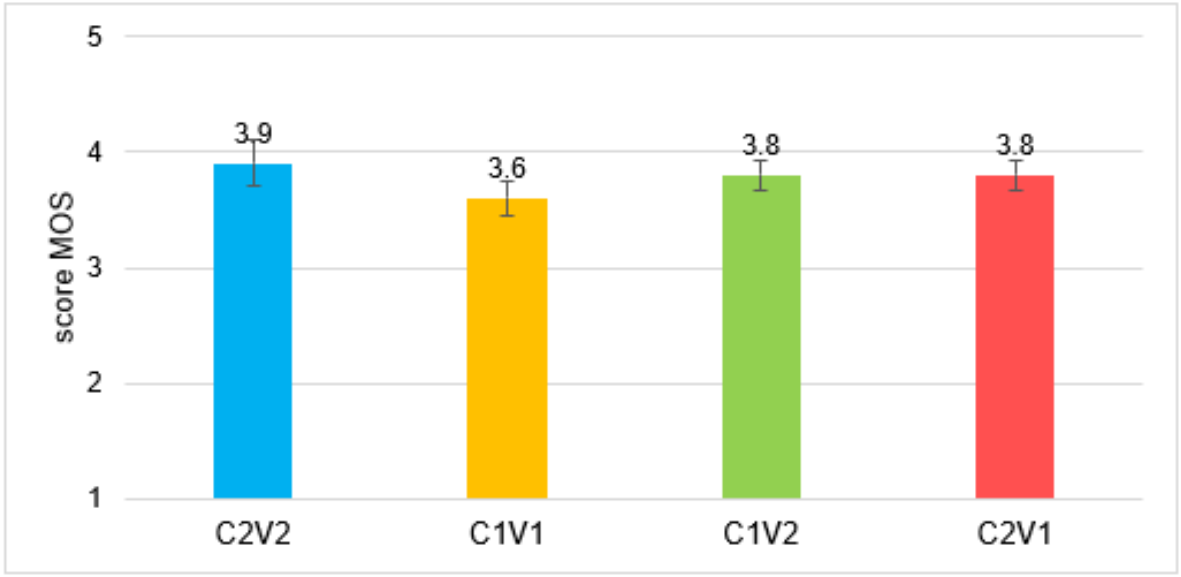
1	2	3	4	5	6	7
Beaucoup plus mauvaise	Plus mauvaise	Un peu plus mauvaise	A peu près la meme	Un peu meilleure	Meilleure	Bien meilleure

Suivant

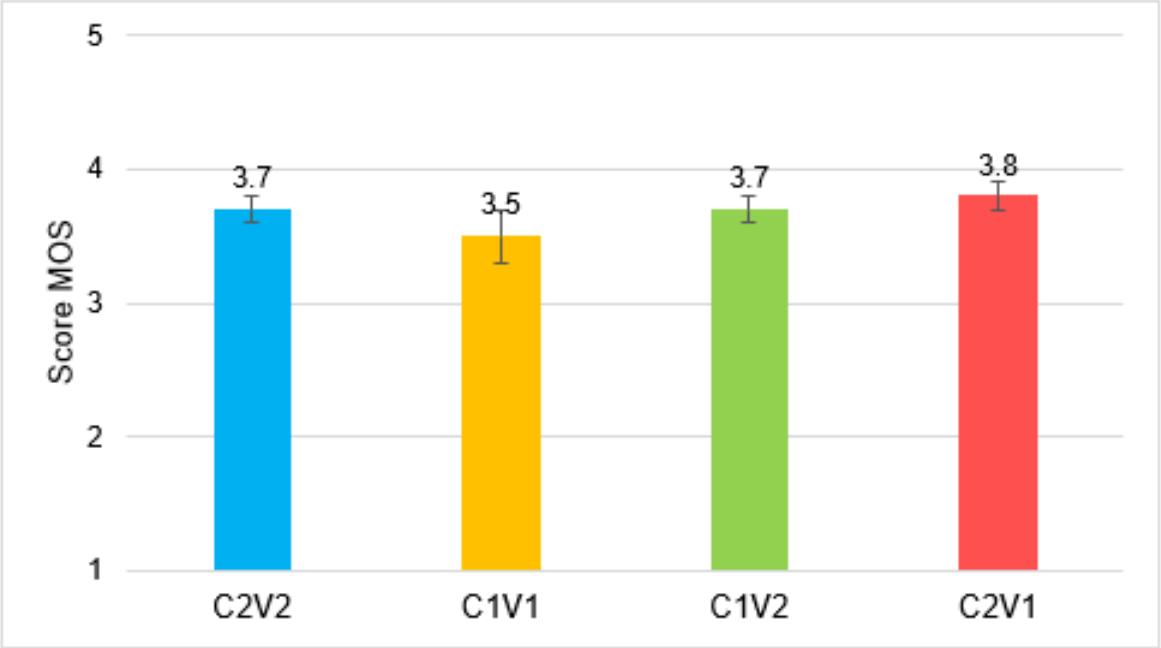
Annexe G

Tests préliminaires

Dans cette première partie des évaluations perceptives, la qualité globale a été évaluée séparément de l'aspect naturel. Ainsi, chaque participant a été mené à répondre à ces deux questions pour chaque signal de parole : *Comment jugez-vous globalement la qualité de ce que vous venez d'entendre ?* et *Tenant compte de l'aspect naturel, comment jugez-vous la qualité de ce que vous venez d'entendre ?* Les résultats obtenus sont représentés comme suit ; la première figure correspond aux scores obtenus après évaluation de la qualité globale. La deuxième figure correspond aux scores d'évaluation de l'aspect naturel.



Évaluation de la qualité globale



Évaluation de l'aspect naturel

Les scores représentés sont associés à leur intervalle de confiance de 95%. Les résultats obtenus pour les deux critères (qualité globale et aspect naturel) sont très similaires. Ceci est dû au fait que les participants étaient des étudiants qui ne sont pas des experts de la synthèse de parole, ainsi ce n'était pas évident de faire la différence entre la qualité globale et l'aspect naturel.

Bibliographie

Abdel-Hamid, O., Abdou, S. M. et Rashwan, M. (2006). Improving Arabic HMM based speech synthesis quality. *In Interspeech, 9th International conference on Spoken Language Processing.*

Abdelmalek, R. et Mnasri, Z. (2016). High quality Arabic text-to-speech synthesis using unit selection. *In SSD, 13th International Multi-Conference on Systems, Signals & Devices*, pages 1–5. IEEE.

Ahmed, B. (2004). Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones.

Al-Ani, S. H. (1970). *Arabic phonology.*

Al-Dakkak, O., Ghneim, N., Zliekha, M. A. et Al-Moubayed, S. (2005). Emotion inclusion in an Arabic text-to-speech. *In EUSIPCO, 13th European Signal Processing Conference*, pages 1–4. IEEE.

Azmy, W. M., Abdou, S. et Shoman, M. (2013). Arabic unit selection emotional speech synthesis using blending data approach. *International Journal of Computer Applications*, 81(8).

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

Banos, E., Erro, D., Bonafonte, A. et Moreno, A. (2008). Flexible harmonic/stochastic modeling for HMM-based speech synthesis. *Jornadas en Tecnologias del Habla*, pages 145–148.

Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S. et Jun, S.-A. (2005). Prosodic typology-the phonology of intonation and phrasing.

Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J. et Keyhl, M. (2013). Perceptual objective listening quality assessment POLQA,

the third generation ITU-T standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384.

Bengio, Y. *et al.* (2009). Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2(1):1–127.

Bennett, C. L. (2005). Large scale evaluation of corpus-based synthesizers : Results and lessons from the blizzard challenge 2005. *In Ninth European Conference on Speech Communication and Technology*.

Benoīt, C., Grice, M. et Hazan, V. (1996). The SUS test : A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18(4):381–392.

Black, A., Taylor, P. et Caley, R. (1998). The festival speech synthesis system.

Black, A. W., Taylor, P. et Caley, R. (2002). The festival speech synthesis system.

Black, A. W., Zen, H. et Tokuda, K. (2007). Statistical parametric speech synthesis. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–1229. IEEE.

Carlson, R. (1995). Models of speech synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 92(22):9932–9937.

Cernak, M. et Rusko, M. (2005). An evaluation of synthetic speech using the PESQ measure. *In European Congress on Acoustics*, pages 2725–2728.

Chai, T. et Draxler, R. R. (2014). Root mean square error RMSE or mean absolute error MAE ? *Geoscientific Model Development Discussions*, 7:1525–1534.

Chalamandaris, A., Tsiakoulis, P., Karabetsos, S. et Raptis, S. (2013). The ILSP/innoetics text-to-speech system for the blizzard challenge 2013. *In Blizzard Challenge Workshop*. Citeseer.

Chenfour, N., Benabbou, A. et Mouradi, A. (2000). Etude et évaluation de la di-syllabe comme unité acoustique pour le système de synthèse arabe PARADIS. *In LREC, The 2nd International Conference on Language Resources and Evaluation*.

Chenfour, N., Mouradi, A. et Benabbou, A. (1997). Synthèse de la parole arabe par concaténation de di-syllabes. *Journées Sientifiques et techniques du réseau Francophone de l Ingénierie de la langue de l'AUPELF-UREF. Avignon France*, pages 459–462.

Chevelu, J., Lolive, D., Maguer, S. L. et Guennec, D. (2015). How to compare TTS systems : A new subjective evaluation methodology focused on differences. *In Sixteenth Annual Conference of the International Speech Communication Association*.

Demri, L. (2016). *Contribution à l'élaboration d'un système de synthèse par concaténation de la parole expressive*. Thèse de doctorat.

Divenyi, P., Greenberg, S. et Meyer, G. (2006). *Dynamics of speech production and perception*, volume 374. Ios Press.

Dutoit, T. et Pagel, V. (1996). Le projet MBROLA : vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. *Actes des Journées d'Etudes sur la parole, Avignon*, pages 441–444.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. et Van der Vrecken, O. (1996). The MBROLA project : Towards a set of high quality speech synthesizers free of use for non commercial purposes. *In ICSLP, Fourth International Conference on Spoken Language,*, volume 3, pages 1393–1396. IEEE.

En-Najjary, T. (2005). *Conversion de voix pour la synthèse de la parole*. Thèse de doctorat, Université Rennes 1.

Engwall, O. (1999). Vocal tract modeling in 3D. *KTH TMH-QPSR*, pages 1–2.

Erro, D., Sainz, I., Luengo, I., Odriozola, I., Sánchez, J., Saratxaga, I., Navas, E. et Hernáez, I. (2010). HMM-based speech synthesis in Basque language using HTS. *Proc. FALA*, pages 67–70.

Fan, Y., Qian, Y., Xie, F.-L. et Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks. *In Fifteenth Annual Conference of the International Speech Communication Association*.

Fant, G. (2012). *Acoustic theory of speech production : with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter.

Fujinaga, K., Nakai, M., Shimodaira, H. et Sagayama, S. (2001). Multiple-regression hidden Markov model. *In ICASSP, International Conference on Acoustics, Speech, and Signal*, volume 1, pages 513–516. IEEE.

Fukada, T., Tokuda, K., Kobayashi, T. et Imai, S. (1992). An adaptive algorithm for Mel-cepstral analysis of speech. *In ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 13–140. IEEE.

Gauvain, J.-L. et Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE transactions on speech and audio processing*, 2(2):291–298.

Gurney, K. (2014). *An introduction to neural networks*. CRC press.

Halabi, N. (2015). *Modern Standard Arabic Speech Corpus*. Thèse de doctorat, University of Southampton.

Halpern, J. *et al.* (2009). Word stress and vowel neutralization in modern standard Arabic. *In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt*.

Houidhek, A., Colotte, V., Mnasri, Z. et Jouvet, D. (2018a). DNN-based speech synthesis for Arabic : Modelling and evaluation. *In International Conference on Statistical Language and Speech Processing*, pages 9–20. Springer.

Houidhek, A., Colotte, V., Mnasri, Z. et Jouvet, D. (2018b). Evaluation of speech unit modelling for HMM-based speech synthesis for Arabic. *International Journal of Speech Technology*, 21(4):895–906.

Houidhek, A., Colotte, V., Mnasri, Z., Jouvet, D. et Zangar, I. (2017). Statistical modelling of speech units in HMM-based speech synthesis for Arabic. *In LTC'2017-8th Language & Technology Conference*.

Hunt, A. J. et Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *In ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 373–376. IEEE.

ITU (1996). 800, methods for subjective determination of transmission quality. *International Telecommunication Union*.

Jurafsky, D. et James, H. (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

Kamina, P. (2014). *Carnet d'anatomie Tome 2–Tête, cou, dos*. MALOINE.

Kawahara, H., Estill, J. et Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. *In Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*.

Kawahara, H., Masuda-Katsuse, I. et De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds. *Speech communication*, 27(3):187–207.

Khalil, K. M. et Adnan, C. (2013). Arabic HMM-based speech synthesis. *In ICEESA, International Conference on Electrical Engineering and Software Applications*, pages 1–5. IEEE.

Khouja, M. K. et Zrigui, M. (2005). Durée des consonnes géminées en parole arabe : mesures et comparaison.

Kim, S.-J., Kim, J.-J. et Hahn, M. (2006). HMM-based Korean speech synthesis system for hand-held devices. *IEEE Transactions on Consumer Electronics*, 52(4).

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3):971–995.

Koulouchli, D. (1976). Contribution à l'étude de l'accent en arabe littéraire. *In Annales de l'Université d'Abidjan, Série H : Linguistique Abidjan*, volume 9, pages 115–130.

Kouloughli, D. (2007). Sur la valeur du tanwin. nouvelle contribution à l'étude du système déterminatif de l'arabe. *Arabica*, 54(1):94.

Krstulovic, S., Hunecke, A. et Schröder, M. (2007). An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. *In INTERSPEECH*, pages 1897–1900. Citeseer.

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. *In IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, volume 1, pages 125–128. IEEE.

Kuhn, R., Junqua, J.-C., Nguyen, P. et Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707.

Labrune, L. (2006). *La phonologie du japonais*, volume 90. Peeters Publishers.

Le Maguer, S., Barbot, N., Boeffard, O. *et al.* (2013). Evaluation of contextual descriptors for HMM-based speech synthesis in French. *In SSW*, pages 153–158.

Leggetter, C. J. et Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech & language*, 9(2):171–185.

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29–45.

Lucci, V. (1983). Etude phonétique du français contemporain à travers la variation situationnelle (débit, rythme, accent, intonation, e muet, liaisons, phonèmes). *Publications de l'Université des Langues et Lettres de Grenoble Grenoble*, pages 1–360.

Lundgren, A. (2005). *An HMM-based text-to-speech system applied to Swedish*. Thèse de doctorat, Royal Institute of Technology (KTH).

Maia, R. d. S., Zen, H., Tokuda, K., Kitamura, T. et Resende Jr, F. G. V. (2003). Towards the development of a Brazilian Portuguese text-to-speech system based on HMM. *In Eighth European Conference on Speech Communication and Technology*.

Mason, A. (2002). The MUSHRA audio subjective test method. *BBC R&D White Paper WHP*, 38.

Masuko, T., Kobayashi, T. et Miyanaga, K. (2004). A style control technique for HMM-based speech synthesis. *In Eighth International Conference on Spoken Language Processing*.

Morise, M. (2012). Platinum : A method to extract excitation signals for voice synthesis system. *Acoustical Science and Technology*, 33(2):123–125.

Morise, M. (2015a). Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67:1–7.

Morise, M. (2015b). Error evaluation of an F0-adaptive spectral envelope estimator in robustness against the additive noise and F0 error. *IEICE transactions on information and systems*, 98(7):1405–1408.

Morise, M., Kawahara, H. et Katayose, H. (2009). Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. *In Audio Engineering Society Conference : 35th International Conference : Audio for Games*. Audio Engineering Society.

Morise, M., Kawahara, H. et Nishiura, T. (2010). Rapid F0 estimation for high-SNR speech based on fundamental component extraction. *Trans. IEICEJ*, 93:109–117.

Morise, M., Yokomori, F. et Ozawa, K. (2016). WORLD : a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884.

Moulines, E. et Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467.

Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J., Le Faucheur, L., Marty, F., Charpentier, F. et Sorin, C. (1990). A real-time French text-to-speech system generating high-quality synthetic speech. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, pages 309–312. IEEE.

Newman, D. (2002). The phonetic status of Arabic within the world's languages : the uniqueness of the lughat al-daad. *Antwerp papers in linguistics.*, 100:65–75.

Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.

Port, R. (2008). ToBI intonation transcription summary. *URL : http ://www. cs. indiana. edu/˜ port/teach/306/tobi. summary. html (visited on//1999)*.

Rajouani, A., Najim, M., Chiadmi, D. et Zyoute, M. (1987). Synthesis-by-rule of Arabic language. *In European Conference on Speech Technology*.

Rebai, I. et BenAyed, Y. (2015). Text-to-speech synthesis system with Arabic diacritic recognition system. *Computer Speech & Language*, 34(1):43–60.

Recommendation, I. (2001). Perceptual evaluation of speech quality PESQ : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.

Rix, A. W., Beerends, J. G., Hollier, M. P. et Hekstra, A. P. (2001). Perceptual evaluation of speech quality PESQ-a new method for speech quality assessment of telephone networks and codecs. *In ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752. IEEE.

Schoeffler, M., Stöter, F.-R., Edler, B. et Herre, J. (2015). Towards the next generation of web-based experiments : A case study assessing basic audio quality following the ITU-R recommendation bs. 1534 MUSHRA. *In 1st Web Audio Conference*, pages 1–6.

Schroeter, J. (2004). Fifty years of progress in speech synthesis. *The Journal of the Acoustical Society of America*, 116(4):2497–2497.

Sejnowski, T. (1987). Nettalk : A parallel network that learns to read aloud. *Complex Systems*, 1:145–168.

Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. et Kitamura, T. (2002). Eigenvoices for HMM-based speech synthesis. *In Seventh International Conference on Spoken Language Processing*.

Shinoda, K. et Watanabe, T. (2001). MDl-based context-dependent subword modeling for speech recognition. *Acoustical Science and Technology*, 21(2):79–86.

Silén, H., Helander, E., Nurminen, J. et Gabbouj, M. (2010). Analysis of duration prediction accuracy in HMM-based speech synthesis. *In Speech Prosody 2010-Fifth International Conference*.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. et Hirschberg, J. (1992). ToBI : A standard for labeling English prosody. *In Second International Conference on Spoken Language Processing*.

Simon, A. C., Avanzi, M. et Goldman, J.-P. (2008). La détection des proéminences syllabiques. un aller-retour entre l'annotation manuelle et le traitement automatique. *In Congrès mondial de linguistique française*, page 151. EDP Sciences.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M. et Richards, C. (2001). Normalization of non-standard words. *Computer speech & language*, 15(3):287–333.

Sydeserff, H., Caley, R., Isard, S. D., Jack, M. A., Monaghan, A. I. et Verhoeven, J. (1992). Evaluation of speech synthesis techniques in a comprehension task. *Speech Communication*, 11(2-3):189–194.

Tachibana, M., Izawa, S., Nose, T. et Kobayashi, T. (2008). Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, pages 4633–4636. IEEE.

Tachibana, M., Yamagishi, J., Masuko, T. et Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE transactions on information and systems*, 88(11):2484–2491.

Tamura, M., Masuko, T., Tokuda, K. et Kobayashi, T. (2001). Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 805–808. IEEE.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *The Journal of the acoustical society of America*, 107(3):1697–1714.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Toda, T. et Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, 90(5):816–824.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. et Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1315–1318. IEEE.

Tokuda, K., Zen, H. et Black, A. W. (2002). An HMM-based speech synthesis system applied to English. *In IEEE Speech Synthesis Workshop*, pages 227–230.

Vilain, C. E. (2002). *Contribution à la synthèse de parole par modèle physique. Application à l'étude des voix pathologiques*. Thèse de doctorat, Institut National Polytechnique de Grenoble-INPG.

Watson, J. C. (2002). *The phonology and morphology of Arabic*. Oxford University Press on Demand.

Watts, O., Henter, G. E., Merritt, T., Wu, Z. et King, S. (2016). From HMMs to DNNs : where do the improvements come from ? *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, pages 5505–5509. IEEE.

Wester, M., Valentini-Botinhao, C. et Henter, G. E. (2015). Are we using enough listeners ? no !—an empirically-supported critique of interspeech 2014 TTS evaluations. *In Sixteenth Annual Conference of the International Speech Communication Association*.

Wu, C.-S. et Hsieh, Y.-F. (2000). Articulatory speech synthesizer. *In Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, pages 345–352.

Wu, Z., Watts, O. et King, S. (2016). Merlin : An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*.

Yamagishi, J. et Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE TRANSACTIONS on Information and Systems*, 90(2):533–543.

Yamagishi, J., Masuko, T. et Kobayashi, T. (2004). HMM-based expressive speech synthesis-towards TTS with arbitrary speaking styles and emotions. *In Proc. of Special Workshop in Maui (SWIM)*.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. et Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. *In Fifth International Conference on Spoken Language Processing*.

Young, S. J. et Young, S. (1993). *The HTK hidden Markov model toolkit : Design and philosophy*. University of Cambridge, Department of Engineering.

Zaki, A., Rajouani, A. et Najim, M. (2001). Synthesizing intonation of standard Arabic language. *In 7th European Conference on Speech Communication and Technology*.

Zen, H. (2006). An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, 133.

Zen, H. (2013). Deep learning in speech synthesis. *In SSW*, page 309.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W. et Tokuda, K. (2007). The HMM-based speech synthesis system HTS version 2.0. *In SSW*, pages 294–299. Citeseer.

Zen, H. et Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, pages 4470–4474. IEEE.

Zen, H. et Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, pages 3844–3848. IEEE.

Zen, H., Senior, A. et Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. *In ICASSP, International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966. IEEE.

Zen, H. et Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for blizzard challenge 2005.

Zen, H., Tokuda, K. et Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.