الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR- ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR - ANNABA

جامعة باجي مختار - عنابة

Faculté des sciences de l'ingéniorat
Département d'électronique

Année : 2018

## Thèse

**Présentée en vue de l'obtention du diplôme de Doctorat en Sciences**

Spécialité : Électronique.

## Intitulé

# Représentations paramétriques du signal de la parole : Application à la langue Arabe

Présentée par : **TABET Youcef**

Dirigée par : **BOUGHAZI Mohamed** **Pr. Université de Annaba**

Co-dirigée par : **LAFIFI Saddek** **Pr. Université de Annaba**

Devant le jury

Président : **LARBI Allel** **Pr. Université de Annaba**

Examinateurs

**BENNACER Layachi** **Pr. Université de Guelma**

**LACHOURI Abderrezak** **Pr. Université de Skikda**

**BOUROUBA Houcine** **MCA. Université de Guelma**

## ملخص بالعربية

في سياق تركيب الكلام ومن أجل الحفاظ على جودة عالية ، قد يكون من المفيد تمثيل إشارات الكلام من خلال نماذج أو تمثيلات . وقد تم بالفعل عرض تمثيلات مختلفة للكلام، وستناقش التمثيلات الأكثر شعبية في هذه الرسالة. سيتم التركيز على التمثيل الجيبي التكيفي ، لأنه يبدو نموذجًا واعدًا للكلام. الهدف هنا ، سيكون تمثيل إشارة الكلام من خلال نموذج بسيط نسبيا ،مرن، عالي الجودة وقوي في إعادة تركيب الكلام .

بناء ا على الاداء المميز للتمثيلات الجيبية التكيفية المقترحة حديثا تم عرض نموذج جديد لتركيب الكلام في هده الرسالة يسمي بالتمثيل الجيبي التكيفي المعدل حيث اقترحت فيه تحسينات كبيرة في كل من مرحلة التحليل والتكيف. أولا سيتم استخدام تمثيل شبه توافقي في مرحلة التحليل من اجل الحصول على تقدير أولي للمعلمات الآنية للنموذج المقترح. بعد دلك في المرحلة التكيفية استخدمت خطة تكيفية مقترنة بآلية تصحيح التردد من أجل الحصول على معلمات أكثر دقة للنموذج المقترح. وأخيرا لإعادة تركيب الكلام، تجمع كل مكونات النموذج الآنية.

أثبتت اختبارات تقييمية ان النموذج الجديد الجيبي التكيفي المعدل عالي الجودة عند تطبيقه في تمثيل اشارات الكلام التوافقية ايضا تم الحصول على جودة عالية وكلام مركب شفاف شبيه بالكلام الأصلي وفقا للنتائج التي تم الحصول عليها من اختبارات الاستماع .

# Abstract

In the context of speech synthesis and in order to maintain high speech quality, it may be advantageous to encode speech signals by speech signal representations or models. Various speech representations have already been proposed in the literature and the more popular ones are discussed in turn in this thesis. Emphasis will be given in adaptive sinusoidal representation, since it seems to be more promising and robust model of speech. It would be desirable a speech signal representation, that is relatively simple, flexible, high quality, and robust in resynthesis.

Based on the performance of the recently suggested adaptive Sinusoidal Models (aSMs) of speech,a refined adaptive sinusoidal representation of speech is proposed in this thesis. This model is referred to as Refined adaptive Sinusoidal Representation (R_aSR). Significant refinements are proposed at both the analysis and adaptive stages. First, a quasi-harmonic representation of speech is used in the analysis stage in order to obtain an initial estimation of the instantaneous model parameters. Next, in the adaptive stage, an adaptive scheme combined with an iterative frequency correction mechanism is used to allow a robust estimation of model parameters (amplitudes, frequencies, and phases). Finally, the speech signal is reconstructed as a sum of its estimated time-varying instantaneous components after an interpolation scheme.

Objective evaluation results prove that the suggested R_aSR achieves high quality reconstruction when applied in modeling voiced speech signals compared to state-of-the-art models. Moreover, transparent perceived quality was attained using the R_aSR according to results obtained from listening evaluation tests.

Keywords: Speech representation, Speech analysis, Speech synthesis, Adaptive sinusoidal modeling.

# Résumé

Dans le contexte de la synthèse vocale et afin de maintenir une qualité de parole élevée, il peut être avantageux de coder les signaux de la parole par des représentations ou des modèles de signaux vocaux. Diverses représentations de la parole ont déjà été proposées dans la littérature et les plus populaires d'entre elles seront discutées à tour de rôle dans cette thèse. L'accent sera mis sur la représentation sinusoïdale adaptative, car elle semble être un modèle prometteur et robuste. L'objectif est d'avoir une représentation du signal de la parole, qui soit relativement simple, flexible, de haute qualité et robuste en resynthèse.

Tenant compte des performances obtenues par les modèles sinusoïdaux adaptatives (aSMs) récemment suggérés, une représentation sinusoïdale adaptative raffinée de la parole est proposée dans cette thèse. Cette représentation est appelée Représentation Sinusoïdale adaptative Raffinée (R_aSR). Des améliorations significatives sont proposées aux étapes d'analyse et d'adaptation. Tout d'abord, une représentation quasi-harmonique de la parole est utilisée dans la phase d'analyse afin d'obtenir une estimation initiale des paramètres instantanés du modèle. Ensuite, dans l'étape d'adaptation, un schéma adaptatif combiné avec un mécanisme itératif de correction de la fréquence fondamentale est utilisé pour permettre une estimation robuste des paramètres du modèle (amplitudes, fréquences et phases). Enfin, le signal vocal est reconstruit en tant que somme de ses composantes instantanées après un mécanisme d'interpolation.

Les résultats des évaluations objectives prouvent que la représentation suggérée réalise une reconstruction de haute qualité lorsqu'elle est appliquée à la modélisation de signaux vocaux voisés. De plus, d'après les résultats des évaluations subjectives, la qualité perçue du signal vocal reconstruit était naturelle et transparente.

Mots cl®s : Repr®sentation de la parole, Analyse de la parole, Synth¯se de la parole, Mod®lisation sinusoµdale adaptative.

*A mes parents*

*A ma femme*

*A mes enfants*

# Remerciements

Je tiens à remercier tout d'abord mon directeur de thèse le professeur. **BOUGHAZI Mohamed** de m'avoir donné l'opportunité de faire cette thèse et de m'avoir apporté un soutien précieux tout au long de mon travail de recherche.

Un grand merci également à mon co-directeur de thèse le professeur **LAFIFI Saddek** pour sa disponibilité et son aide.

Je tiens à remercier les membres du jury qui m'ont fait l'honneur de bien vouloir évaluer mon travail. D'abord, je remercie monsieur. **LARBI Allel**, professeur à l'université de Annaba, pour l'honneur qu'il m'a fait, en acceptant la présidence de ce jury. Je remercie également monsieur **BENNACER Layachi**, professeur à l'université de Guelma, monsieur **LACHOURI Abderrezak** , professeur à l'université de Skikda et monsieur **BOUROUBA Houcine**, MCA à l'université de Guelma, d'avoir accepté de rapporter mon travail.

Et puis, bien sûr, je n'oublie pas de remercier tous les enseignants et le personnel du département d'électronique.

Merci à tous ceux qui m'ont aidé et soutenu pendant ces années de recherches.

# Liste des publications et des communications

Majeures parties des chapitres 1 et 2 ont été publiées dans

— **"7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA)- indéxé dans IEEE Explorer 2011"** en tant que :

Y.Tabet, M.Boughazi . **Speech Synthesis Techniques. A Survey**. 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), p67-70, (2011).

— **"Procedia Computer Science, ELSEVIER, ScienceDirect, 2015"** en tant que :

Y. Tabet, M. Boughazi, S. Affifi. **A Tutorial on Speech Synthesis Models**. The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT). 73 48-55, (2015).

Majeures parties des chapitres 3, 4 et 5 ont été publiées dans

— **"International Journal of Speech Technology (Springer 2018)"** en tant que :

Y. Tabet, M. Boughazi, S. Afifi . **Speech Analysis and Synthesis with a Refined Adaptive Sinusoidal Representation**. International Journal of Speech Technology (IJST). (2018).

# Liste des acronymes

| **Acronyme** | Signification |
|---|---|
| **ABS** | Analysis By Synthesis (Analyse Par Synthèse) |
| **aHM** | adaptif Harmonic Model (Modèle Harmonique adaptatif) |
| **aHNM** | adaptif Harmonic plus Noise Model (Modèle Harmonique adaptatif plus Bruit) |
| **AIR** | Adaptive Iterative Refinement (Raffinement Itératif Adaptatif) |
| **aQHM** | adaptif Quasi Harmonic Model (Modèle Quasi Harmonique adaptatif) |
| **aQHNM** | adaptif Quasi Harmonic plus Noise Model (Modèle Quasi Harmonique adaptatif plus Bruit) |
| **AR** | Autorégressif |
| **aSMs** | adaptif Sinusoidal Models (Modèles adaptatifs Sinusoïdaux) |
| **eaQHM** | extended adaptif Quasi Harmonic Model (Modèle Quasi Harmonique adaptatif étendu) |
| **eaQHNM** | extended adaptif Quasi Harmonic plus Noise Model (Modèle Quasi Harmonique adaptatif étendu plus Bruit) |
| **FFT** | Fast Fourier Transform (Transformée de Fourier Rapide) |
| **HNM** | Harmonic plus Noise model (Modèle Harmonique plus Bruit) |
| **MBROLA** | Multi-Band Re-synthesis Overlap-Add (Re-synthèse Multi-Bande par Chevauchement-Addition |
| **MDCT** | Modified Discrete Cosinus Transformation (Transformation de Cosinus Discrète Modifiée) |
| **MC** | Moindres Carrés |
| **MFCC** | Mel Frequency Cepstral Coefficient (Coefficient Cepstral de Fréquence de Mel) |
| **MOS** | Mean Opinion Score (Score Moyen d'Opinion) |
| **OLA** | Overlap and Add (Chevauchement-Addition) |
| **PSOLA** | Pitch-Synchronous Overlap-Add (Chevauchement-Addition Synchrone avec le Pitch) |
| **QHM** | Quasi Harmonic Model (Modèle Quasi Harmonique) |
| **R_SF** | Représentation Source-Filtre |
| **R_PL** | Représentation Prédictive Linéaire |

**R_S** Représentation Sinusoïdale

**R_DR** Représentation Déterministique plus Résiduel

**R_HB** Représentation Harmonique plus Bruit

**R_DTS** Représentation Déterministique, Transitoire et Stochastique

**SM** modèle Sinusoïdal

**SRER** Signal-to-Reconstruction-Error Ratio (Rapport Signal-Reconstruction-Erreur)

**SWIPE** Sawtooth Waveform Inspired Pitch Estimator (Estimateur de Pitch Inspiré en Forme de Dents de Scie)

**TTS** Text to Speech System (Système de Conversion Texte-Parole)

**TFCT** Transformée de Fourier à Court Terme

**TF** Transformée de Fourier

**TFD** Transformée de Fourier Discrète

**Tz** Transformée en z

**WSOLA** Weighted Synchronized Overlap-Add ( Chevauchement-Addition Pondéré Synchronisé )

# Liste des tableaux

# Table des figures

# Table des matières

# Introduction générale

Dans notre société où la rapidité et l'efficacité sont des qualités clés, une interaction homme-machine via la parole est d'une grande importance. Une telle interaction implique la reconnaissance et la synthèse de la parole. La reconnaissance de la parole consiste à extraire les informations d'un message du signal vocal de manière à contrôler les actions d'une machine en réponse à des commandes parlées. La synthèse de la parole est le processus de créer une réplique synthétique d'un signal de la parole de manière à transmettre un message d'une machine à une personne, dans le but de transmettre l'information dans le message [1].

En synthèse vocale, le but est d'obtenir un signal vocal synthétique non seulement facilement compréhensible, mais aussi indiscernable de celui produit par un être humain, en d'autres termes, créer un système de synthèse vocale possédant des performances proches ou égales aux performances humaines. Par conséquent, au cours des dernières décennies, la parole synthétique a été développée d'une manière régulière afin d'améliorer l'intelligibilité et le caractère naturel de la sortie d'un système de synthèse vocale.

Pour atteindre une synthèse vocale de haute qualité, l'étude des variations temporelles et spectrales des signaux de la parole est d'une grande importance car elles transmettent des informations telles que les mots, l'intention, l'expression, l'intonation, l'accent, l'identité du locuteur, le genre, le style de parler, l'état de santé du locuteur et l'émotion [2]. L'évolution temporelle ou spectrale du signal vocal peut être représentée par un modèle mathématique. Les avantages de l'utilisation d'un modèle sont sa capacité à réduire la redondance du signal acoustique et à définir les paramètres les mieux adaptés au traitement acoustique du signal de la parole [3].

La représentation ou la modélisation du signal vocal joue un rôle important dans plusieurs

applications du traitement de la parole, y compris le codage, l'analyse / synthèse et la reconnaissance de la parole. Dans les systèmes d'analyse / synthèse de la parole, par exemple, un ensemble de paramètres du modèle sont extraits au stade de l'analyse, puis ces paramètres seront utilisés au stade de la synthèse pour reconstruire le signal synthétique. Il serait donc souhaitable d'avoir un modèle paramétrique de la parole, qui soit relativement simple, flexible, de haute qualité et robuste en resynthèse. Par conséquent, un choix approprié du modèle et une estimation précise des paramètres du modèle sont deux éléments clés pour le succès dans toutes les applications de traitement de la parole [4].

Dans le contexte des applications de la synthèse vocale et afin de maintenir une bonne qualité de la parole reconstruite, il peut être avantageux de coder les signaux de la parole par des représentations mathématiques [5, 6]. Dans les systèmes actuels de la synthèse vocale, plusieurs techniques de traitement du signal pour la représentation de la parole ont été développées dans le but de générer une parole sonore naturelle [7]. Par conséquent, une grande variété de représentations de signaux vocaux a été discutée dans la littérature [4, 8]. Parmi elles : la représentation temporelle, la représentation spectrale, la représentation prédictive linéaire, la représentation cepstrale ou homomorphique, la représentation sinusoïdale, etc.

Le modèle de la production de la parole [9] et le modèle sinusoïdal [10] sont les deux principaux modèles utilisés dans la synthèse de la parole. Le premier est un modèle avec plusieurs systèmes en série qui représentent les différentes étapes de la production de la parole humaine. Le deuxième modèle décompose le signal observé en une somme de composantes sinusoïdales, c'est-à-dire une somme de cosinus modulés en fréquence et / ou en amplitude. Dans une application de synthèse de la parole, la sélection d'un modèle dépend de nombreux facteurs tels que la qualité de la parole synthétisée, la facilité d'extraction des paramètres, la modification des paramètres, le nombre de paramètres et la charge de calcul. Des améliorations sur chaque représentation de base (modèle de la production de la parole ou modèle sinusoïdale) ont été proposées au cours des années afin d'obtenir une meilleure qualité du signal de parole reconstruit, les plus populaires d'entre elles seront décrites dans cette thèse.

# Problématique et Contributions

Plusieurs représentations du signal de la parole utilisées dans des applications d'analyse-synthèse vocale ont été donc proposées dans la littérature [7]. La représentation par prédiction linéaire était parmi les modèles du signal de la parole les plus puissants et a été appliquée avec succès dans l'analyse et la synthèse de la parole [11]. Les principaux avantages de l'approche prédictive linéaire sont : la simplicité, la rapidité et le nombre limité de ces paramètres. Cependant, en raison de la nature paramétrique de cette représentation et la simplicité de la modélisation de l'excitation prédictive linéaire, la qualité de la parole produite par les systèmes prédictifs linéaires est dégradée et manque de naturel. Pour pallier ce problème, la représentation par prédiction linéaire a cédé la place à des modèles plus complexes offrant une meilleure qualité de signal. Par exemple, les représentations sinusoïdales [10] sont des représentations assez générales de la parole et peuvent être utilisées dans une large gamme de sons. Elles ont été appliquées avec succès dans l'analyse et la synthèse de la parole. Parce que la représentation sinusoïdale est bien adaptée pour modéliser les phénomènes quasi périodiques qui se produisent typiquement dans les sons voisés, la contrepartie non voisée est mal représentée par ce type de représentation. Pour faire face à ce problème, il a été proposé de décomposer la représentation du signal vocal en deux composantes distinctes (composante sinusoïdale et composante de bruit) [12]. Ce type de représentation hybride améliore énormément la qualité du signal synthétique et elle a été utilisée avec succès dans l'analyse et la synthèse de la parole. Cependant, le principal inconvénient de ce type de représentation est la complexité des calculs par rapport aux approches prédictive linéaire et sinusoïdale.

Autre inconvénient majeur des représentations du signal de la parole citées ci-dessus est la sensibilité à l'estimation des fréquences. Une mauvaise estimation des fréquences entraîne des erreurs de reconstruction élevées. Pour résoudre ce problème, il a été suggéré de représenter les signaux de parole par un modèle quasi harmonique (QHM) [13] dont le principal avantage est sa capacité à corriger les erreurs d'estimations de fréquences d'une manière directe.

Le modèle QHM ainsi que les modèles standards du signal de la parole (prédiction linéaire et sinusoïdale) cités ci-dessus tiennent compte de la stationnarité locale du signal vocal dans leurs représentations. C'est-à-dire, leurs paramètres sont supposés constants sur de courts intervalles de

temps. Cependant, le signal de la parole est considéré comme un signal non stationnaire. Pour traiter les caractéristiques non stationnaires des signaux de la parole, des représentations de signaux vocaux améliorées basées sur des modèles sinusoïdaux adaptatifs ont attiré l'attention des chercheurs en raison de leur capacité à adapter leurs paramètres aux caractéristiques locales (phase / amplitude) du signal vocal analysé [14, 15, 16, 17, 18].

Les principaux objectifs de cette thèse sont donc les suivants : l'exploration de plusieurs représentations paramétriques stationnaires et non stationnaires (adaptatives) du signal de la parole et en se basant sur les performances de ces dernières, nous allons proposer une nouvelle représentation du signal vocal qui donnera une meilleure reconstruction du signal quand elle est appliquée à l'analyse-synthèse vocale. Pour valider la performance de notre nouvelle représentation, une comparaison sera faite avec les différentes représentations stationnaires ou adaptatives de l'état de l'art utilisant différents types de bases de données vocales.

Par conséquent, dans cette thèse, l'accent est mis sur les représentations sinusoïdales adaptatives et nos contributions dans le domaine de la représentation du signal de la parole pour application à l'analyse-synthèse vocale sont les suivantes :

— Exploration des différentes approches stationnaires utilisées pour la représentation paramétrique du signal de la parole, à savoir, la représentation prédictive linéaire, la représentation sinusoïdale et ses extensions et la représentation quasi harmonique.

— Exploration des différentes approches adaptatives récemment développées, utilisées pour la représentation du signal de la parole, à savoir, la représentation quasi harmonique adaptative, la représentation quasi harmonique adaptative étendue et la représentation harmonique adaptative.

— Proposition d'une représentation adaptative sinusoïdale raffinée du signal de la parole [19]en se basant sur les performances obtenues par les représentations sinusoïdales adaptatives du signal de la parole. La représentation ainsi suggérée sera capable d'améliorer l'estimation des paramètres du modèle correspondant et par conséquent une haute qualité

de reconstruction du signal de la parole sera obtenue,

— Développement d'un système complet d'analyse-synthèse vocale utilisant notre représentation adaptative sinusoïdale raffinée du signal de la parole,

— Application de notre nouvelle représentation adaptative sinusoïdale raffinée à l'analyse-synthèse du signal vocal voisé (arabe, anglais). L'accent est mis sur le signal vocal arabe, vu que ce dernier n'a pas été l'objet de beaucoup de travaux contrairement au signal vocal anglais.

## Organisation de la thèse

Outre l'**introduction générale**, où nous avons décrit le contexte général de cette recherche ainsi que les grandes lignes de notre contribution, ce rapport de thèse est divisé en deux grandes parties :

**Partie État de l'art** comportant les chapitres suivants :

— **Le chapitre 1** expose des généralités sur les techniques de synthèse utilisées dans les systèmes de synthèse vocale à partir du texte ainsi que les différentes représentations de base du signal de la parole, à savoir, la représentation source-filtre, la représentation linéaire prédictive et la représentation par la transformée de Fourier ;

— **Le chapitre 2** présente un bref aperçu des représentations sinusoïdales stationnaires du signal de la parole, à savoir, les représentations sinusoïdales uniformes et les représentations sinusoïdales hybrides ;

— **Le chapitre 3** fournit une description des représentations sinusoïdales adaptatives du signal vocal récemment suggérées, à savoir, le modèle quasi harmonique, le modèle adaptatif quasi harmonique, le modèle adaptatif quasi harmonique étendu et le modèle adaptatif harmonique ;

**Partie Contributions :** comportant les deux chapitres suivants :

— **Le chapitre 4** est consacré à notre principale contribution qui est la proposition d'une représentation sinusoïdale adaptative raffinée de la parole, ainsi que les étapes d'analyse, d'adaptation et de synthèse qui vont avec ;

— **Le chapitre 5** décrit en détail le système d'analyse-synthèse vocale basé sur notre représentation sinusoïdale adaptative raffinée et son application à deux types de bases de données de parole (Anglaise et Arabe).

Finalement, nous concluons par une récapitulation de la recherche effectuée et nous donnons des perspectives ouvertes de ce travail.

# Première partie

# Partie État de l’art

# Chapitre 1

# Généralités sur les techniques et les modèles de la synthèse vocale

Pour produire un signal de la parole de haute qualité par une machine ou un ordinateur, un système de synthèse de la parole à partir d'un texte (TTS) utilise une technique de synthèse vocale appropriée aux besoins de l'application demandée. Également, chaque technique de synthèse vocale peut utiliser une représentation (ou un modèle) paramétrique du signal de la parole afin d'obtenir une meilleure qualité de la reconstruction vocale.

Dans ce chapitre, nous allons présenter un aperçu du fonctionnement d'un système de synthèse de la parole à partir d'un texte ainsi que les principales techniques de synthèse vocale, à savoir, la synthèse par formants, la synthèse articulatoire, la synthèse concaténative, la synthèse par sélection d'unité et enfin la synthèse par modèle de Markov caché. Nous donnons un bref historique des représentations les plus populaires du signal de la parole. Ensuite, nous décrivons le mécanisme de la production de la parole suivi des différents types de représentations de base du signal de la parole, à savoir, la représentation source-filtre, la représentation prédictive linéaire et la représentation par transformée de Fourier. Des exemples expérimentaux utilisant ces différents types de représentations du signal de la parole sont également présentés dans ce chapitre.

## 1.1 Bref aperçu d'un système de synthèse de la parole à partir du texte (TTS)

La synthèse de la parole à partir du texte consiste en un ensemble de traitements permettant à un ordinateur de transformer un texte écrit en un message vocal. L'objectif est de produire une voix synthétique intelligible et naturel [20]. Le processus de transformation du text écrit en un message vocal s'effectue en trois principales phases. Les deux premières phases permettent le passage de la représentation orthographique du texte à la représentation phonétique munie d'une description prosodique. La dernière phase permet la génération du signal acoustique en utilisant différentes méthodes de synthèses. Ces dernières mettent en œuvre des techniques de traitement du signal avec des représentations du signal de la parole pour atteindre une synthèse parfaite (figure 1.1).

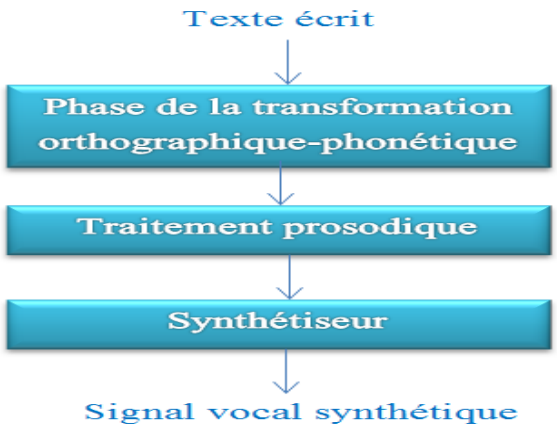


FIGURE 1.1 – Schéma simplifié du système de la synthèse à partir du texte

## 1.2 Bref aperçu des techniques de la synthèse vocale

Il existe dans la literature plusieurs approches du traitement de signal utilisées par les systèmes de la synthèse vocale à partir du texte [21]. Nous présentons par la suite un bref aperçu des plus importantes d'entre elles.

### 1.2.1 Synthèse par formants

La synthèse par formants modélise les fréquences du signal vocal. Les formants sont les fréquences de résonance du conduit vocal. Le signal vocal est synthétisé en utilisant ces fréquences estimées. Dans la synthèse par formants, l'hypothèse de base est que la fonction de transfert de conduit vocal peut être modélisée de façon satisfaisante en simulant les fréquences et les amplitudes des formants. La synthèse consiste donc en la reconstruction artificielle des caractéristiques du formant à produire. Ceci est fait en excitant un ensemble de résonateurs par une source de voisement ou un générateur de bruit pour atteindre le spectre de la parole souhaité, et en contrôlant la source d'excitation pour simuler soit le voisement ou le non voisement. L'ajout d'un ensemble d'anti-résonateurs permet en outre la simulation des effets du tube nasal, des fricatives et des plosives. La spécification d'environ 20 ou plus de tels paramètres peut conduire à une restitution satisfaisante du signal vocal. L'avantage de cette technique est que ses paramètres sont fortement corrélés avec la production et la propagation du son dans le tractus oral. Le principal inconvénient de cette approche est que les techniques automatiques de spécification des paramètres des formants sont encore largement insatisfaisantes, et que, par conséquent, la majorité des paramètres doit encore être optimisée manuellement [22].

La synthèse de formant n'utilise aucun échantillon de parole humain mais s'appuie sur des règles écrites par des linguistes pour générer les paramètres qui permettront la synthèse de parole, et pour faire face à la transition d'un phonème à un autre, c'est-à-dire la coarticulation. Pour écrire les règles, les linguistes ont étudié les spectrogrammes et ont déduit les règles d'évolution des formants. Cependant, nous ne connaissons pas encore la règle optimale pour le faire [23] . De plus, la forme d'onde de la parole est produite naturellement dans un processus si complexe que, actuellement, les règles peuvent uniquement modéliser les caractéristiques de la forme d'onde de la parole. Par conséquent, le signal de parole synthétisé a un effet artificiel et il est perçu comme un son robotique non naturel. Cependant, le signal vocal synthétisé basé sur des règles est très intelligible, même à haute vitesse, ce qui est très utile pour les malvoyants qui naviguent rapidement sur des ordinateurs en utilisant un lecteur d'écran. De plus, lorsque les coûts de mémoire et de traitement sont limités, comme dans les systèmes embarqués, ces synthétiseurs sont plus intéressants parce qu'ils n'ont pas de base de données des échantillons des signaux vocaux.

### 1.2.2 Synthèse articulatoire

La synthèse articulatoire génère la parole par modélisation directe du comportement de l'articulateur humain. Donc, en principe, c'est la méthode la plus satisfaisante pour produire un signal vocal de haute qualité. En pratique, c'est l'une des méthodes les plus difficiles à mettre en œuvre. Les paramètres de contrôle articulatoire sont les suivant : l'ouverture des lèvres, la protrusion des lèvres, la position de la pointe de la langue, la hauteur de la pointe de la langue, la position de la langue et enfin la hauteur de la langue [24].

Il y a deux difficultés dans la synthèse articulatoire. La première difficulté est l'acquisition de données pour le modèle articulatoire. Ces données proviennent généralement de la photographie Xray[23]. Les données de rayons X ne caractérisent pas les masses ou les degrés de liberté des articulateurs. La deuxième difficulté est de trouver un équilibre entre un modèle très précis et un modèle facile à concevoir et à contrôler. En général, les résultats de la synthèse articulatoire ne sont pas aussi bons que les résultats de la synthèse des formants ou des résultats de la synthèse concaténative.

### 1.2.3 Synthèse concaténative

La principale limitation de la synthèse par formants et de la synthèse articulatoire n'est pas tellement la production d'un signal de parole à partir de la représentation paramétrique, mais la difficulté est de trouver ces paramètres à partir de la spécification d'entrée créée par le processus d'analyse de texte. Pour surmonter cette limitation, la synthèse concaténative suit une approche axée sur les données. La synthèse concaténative génère un signal vocal en connectant des unités de parole naturelles pré-enregistrées. Ces unités peuvent être des mots, des syllabes, des demi-syllabes, des phonèmes, des diphones ou des triphones. La longueur de l'unité affecte la qualité du signal vocal synthétisé. Avec des unités plus longues, le naturel augmente, moins de points de concaténation sont nécessaires, mais plus de mémoire est nécessaire et le nombre d'unités stockées dans la base de données devient très important. Avec des unités plus courtes, il y a moins de mémoire, mais les techniques de collecte et de marquage d'échantillons deviennent plus complexes [20] ; Les unités les plus largement utilisées dans la synthèse concaténative sont les diphones. Un diphone est une unité qui commence au milieu d'un phone et s'étend au milieu du suivant. Les

diphones ont l'avantage de modéliser la co-articulation en incluant la transition vers le phone suivant à l'intérieur du diphone lui-même. La liste complète des diphones est appelée inventaire de diphones, et une fois déterminée, les diphones doivent être trouvés dans le vrai signal de la parole. Pour construire l'inventaire de diphones, un discours naturel doit être enregistré tels que tous les phonèmes dans tous les contextes possibles (allophones) sont inclus, puis les diphones doivent être étiquetés et segmentés. Une fois l'inventaire de diphones construit, la hauteur(pitch) et la durée de chaque diphone doivent être modifiés pour correspondre à la partie prosodique de la spécification. D'autre part, la synthèse concaténative produit ainsi de la parole en concaténant de petites unités de discours pré-enregistrées, mais dans le cas où pas seulement une, mais des centaines de réalisations de chaque unité de parole phonétique sont présentes dans un inventaire, un processus de sélection d'unités doit avoir lieu pour créer la séquence d'unités synthétique finale. Une telle méthode de synthèse vocale est également appelée synthèse vocale basée sur le corpus et sera décrite dans la section qui suit.

### 1.2.4 Synthèse par sélection d'unité

Dans la synthèse concaténative, les diphones doivent être modifiés par des méthodes de traitement du signal pour produire la prosodie désirée. Cette modification entraîne des artefacts dans le signal de la parole qui peut rendre le discours non naturel. La synthèse par sélection d'unités (également appelée synthèse concaténative basée sur un corpus) résout ce problème en stockant dans l'inventaire des unités (plusieurs instances de chaque unité avec des prosodies variables) [25]. L'unité qui correspond le mieux à la prosodie cible est sélectionnée et concaténée de sorte que les modifications prosodiques nécessaires sur l'unité sélectionnée sont minimisées. Puisque plusieurs instances de chaque unité sont stockées dans l'inventaire de l'unité, un algorithme de sélection d'unités est nécessaire pour choisir les unités qui correspondent le mieux à la spécification cible. Cette sélection est basée sur la minimisation de deux types de fonctions de coûts (cible et jointure).

Dans le cas de la sélection automatique d'unités, l'influence Co-articulaire n'est pas limitée au dernier phonème. La base de données est beaucoup plus grande (1-10 heures) et comprend plusieurs occurrences de chaque unité acoustique, capturée dans divers contextes (comme ses phonèmes voisins bien sûr, mais aussi sa hauteur, sa durée, sa position dans la syllabe, etc.).

En dehors de ce caractère naturel, les techniques de sélection d'unités présentent plusieurs inconvénients. Elles s'appuient sur une très grande base de données, ce qui implique, d'une part, un temps de développement et un coût considérables pour collecter et étiqueter les données, et d'autre part, un important besoin en ressources de mémoires pour stocker les données. Le deuxième inconvénient est l'étiquetage incorrect et l'apparition de contextes cibles invisibles conduisent à des fragments de signaux vocaux synthétisés de très mauvaise qualité. Ce phénomène de contextes invisibles pourrait bien ne jamais être complètement surmonté avec la synthèse concaténative.

### 1.2.5 Synthèse par modèle de Markov caché (HMM)

La synthèse par modèle de Markov caché (HMM) se compose de deux phases principales, la phase d'entraînement et la phase de synthèse [26]. Lors de la phase d'entraînement, il convient de décider à quelles caractéristiques les modèles doivent être formés. Les coefficients cepstraux de fréquence Mel (MFCC) et leurs dérivées première et seconde sont les types les plus communs de caractéristiques utilisées. Les caractéristiques sont extraites par trame et placées dans un vecteur de caractéristiques. L'algorithme de Baum-Welch est utilisé avec les vecteurs de caractéristiques pour produire des modèles pour chaque phone. Un modèle se compose généralement de trois états qui représentent le début, le milieu et la fin du phone. La phase de synthèse comprend deux étapes : Premièrement, les vecteurs de caractéristiques pour une séquence de phone donnée doivent être estimés. Deuxièmement, un filtre est implémenté pour transformer ces vecteurs caractéristiques en signaux audio. La qualité du signal vocal généré par HMM n'est pas aussi bonne que la qualité de la parole générée à partir de la synthèse par sélection d'unités.

## 1.3 Bref historique des représentations du signal de la parole

Puisque la recherche présentée dans cette thèse sera principalement appliquée au domaine de la représentation du signal de la parole pour l'analyse-synthèse vocale, il est important de discuter des travaux antérieurs dans ces domaines d'application, afin de fournir un cadre historique pour la présente recherche et des points de référence significatifs aux lecteurs. Par conséquent, la section suivante examinera les principales représentations du signal de parole utilisées par les systèmes d'analyse-synthèse vocale.

La représentation par vocodeur inventée par Dudley dans les années trente [27] était la première tentative pour représenter le signal de parole par une source sonore d'excitation (périodique ou bruit) et un filtre vocal (une banque de filtres passe-bande analogiques). Cette représentation a une correspondance directe avec le mécanisme de production de la parole et l'objectif principal était d'obtenir une transmission et un stockage efficaces du signal vocal. Dans les années soixante, une autre technique pour encoder et représenter la parole appelée vocodeur de phase a été suggérée par Flanagran et Golden [28]. L'idée de base derrière cette approche est de représenter la parole en termes de son amplitude et de sa phase à court terme. Après cela, une formulation numérique du vocodeur de phase a été introduite par Portnoff [29] en représentant une forme d'onde de la parole par sa transformée de Fourier à court terme (TFCT). Une efficacité de calcul a été obtenue en utilisant l'algorithme de la transformée de Fourier rapide FFT (Fast Fourier Transform).

À la fin des années cinquante, Fant a développé le fameux modèle linéaire de la production de parole appelé représentation source-système ou bien source-filtre [9]. Dans ce modèle, une source de train d'impulsions périodiques est appliquée à un système linéaire variant lentement dans le temps (modèle du conduit vocal) pour la partie voisée du signal de parole. Cependant, un bruit aléatoire est utilisé pour exciter le système pour une partie non voisée du signal de parole. Le filtre du conduit vocal est supposé être un modèle tout pôle et ses paramètres sont estimés via une analyse basée sur la technique de la prédiction linéaire [8, 30].

Dans les années quatre-vingt, des modèles sinusoïdaux plus complexes que le modèle source-filtre et qui offraient une meilleure qualité de signal sont apparus dans la littérature [31, 32, 33, 10, 34]. Le modèle sinusoïdal est une représentation assez générale qui peut être utilisée dans une large gamme de sons et offre un gain en flexibilité et en qualité perceptive par rapport à d'autres techniques [4].

Parce que le modèle sinusoïdal est adapté à la modélisation des sons harmoniques, les sons non voisés sont mal représentés par ce modèle. Aussi, pour modéliser précisément le bruit, un nombre infini de sinusoïdes sont nécessaires et ce modèle a des limites comme la grande base de données et la complexité de calcul. Pour faire face à ce problème, diverses représentations sinusoïdales hybrides qui décomposent le signal de parole en deux parties distinctes (partie sinusoïdale

et partie bruit) sont apparues [35, 36, 37, 38, 12] et il a été montré que ces modèles améliorent la qualité perçue de la parole reconstruite par rapport aux modèles sinusoïdaux standards.

Récemment, la représentation harmonique plus bruit proposée dans [38] a été réintroduite dans [13] et il a été montré que ce type de représentation est capable de résoudre les problèmes liés à l'estimation de fréquences. Ce qui a produit une meilleure estimation des autres paramètres (amplitudes et phases) de la représentation. La version revisitée du modèle harmonique plus bruit est nommée modèle quasi harmonique (QHM). En se basant sur ce modèle, autres types de représentations nommées modèles sinusoïdaux adaptatifs (aSMs) [14, 15, 16, 17, 18] ont été suggérés pour bien modéliser le caractère non stationnaire du signal de la parole. Il a été montré que la qualité de la représentation et de la reconstruction vocale ont été nettement améliorées par rapport aux autres représentations de l'état de l'art.

## 1.4 Signal de la parole et Mécanisme de production

Le signal de la parole est produit par l'appareil phonatoire humain présenté en figure 1.2 [39]. Trois acteurs principaux contribuent à cette production et qui sont : les poumons, le larynx et le conduit vocal [40]. Une énergie ventilatoire générée par les poumons va être utilisée pour mettre en mouvement les cordes vocales au niveau du larynx pour produire les sons voisées et/ou à générer des bruits de friction ou d'explosion. Ensuite, différent types de sons de la parole peuvent être produit par la réalisation d'une gestuelle articulatoire au niveau du conduit vocal constitué de l'ensemble de cavités se situant entre le larynx et les lèvres (cavité pharyngale, cavité nasale, cavité buccale et cavité labiale) [39, 40].

On peut dire donc que le signal de parole est le résultat de l'excitation des cavités supra glottiques par une ou deux sources acoustiques (source laryngienne, bruits d'explosion ou de friction). Suivant l'hypothèse de quasi-stationarité à court terme du signal de parole, et dans le cadre des méthodes de modélisation des signaux stationnaires, il est possible de poser un modèle qui permet de rendre compte, d'une manière simple et relativement efficace, des interactions à court terme du processus de production de la parole.

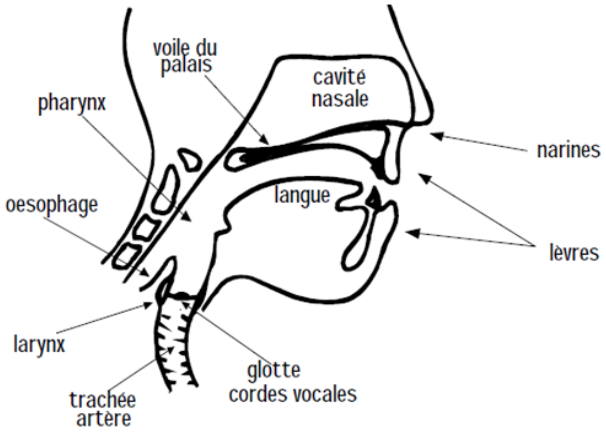


FIGURE 1.2 – Appareil phonatoire

Le système phonatoire produit un signal de parole qui apparaît comme une variation temporelle de la pression d'air. Donc, d'un point physique, on peut considérer que la parole est un phénomène acoustique produit par le système phonatoire. Un ensemble de grandeurs tel que la fréquence fondamentale, l'énergie et le spectre fréquentiel peuvent caractériser un signal de parole.

La fréquence fondamentale ou la période du pitch représente la fréquence de vibration des cordes vocales, caractérisant ainsi que les segments voisés de la parole, à l'intérieur desquels elle évolue lentement dans le temps. Pour les segments non voisés, c'est-à-dire, les segments où il n'y a pas de vibration de cordes vocales, cette fréquence fondamentale est nulle.

L'intensité sonore d'un segment de parole définit l'énergie du signal qui est en générale plus élevée pour les segments de parole voisés que pour les segments non-voisés.

Enfin, le spectre fréquentiel d'un signal de la parole consiste en une représentation du signal de la parole dans le domaine fréquentiel utilisant la transformée de Fourier. Ce spectre peut s'étendre jusqu'à 12kHz, mais on admet généralement que le spectre utile est limité entre 4kHz et 8kHz (le cas d'une transmission téléphonique)[39, 40].

## 1.5 Modélisation de la production de la parole

De nombreuses techniques d'analyse-synthèse de la parole sont basées sur ce qu'on appelle le modèle de production de la parole [8]. Afin de bien modéliser la production du signal de la pa-

role, une grande variété de représentations de signaux vocaux ont été discutées dans la littérature [4]. Parmi elles : la représentation temporelle (c'est-à-dire la forme d'onde de la parole) ; représentation spectrale ; représentation prédictive linéaire, représentation cepstrale ou homomorphique, représentation sinusoïdale, etc. La discussion dans les prochaines sections, se concentre uniquement sur les représentations de signaux vocaux les plus importantes, à savoir, la représentation source-filtre, la représentation linéaire prédictive et enfin la représentation par la transformée de Fourier à court terme.

### 1.5.1 Représentations temporelles du signal de la parole

#### 1.5.1.1 Représentation source-filtre (R_SF)

On peut modéliser la production phonatoire par une représentation R_SF [9] , où le rôle de la source est joué par les cordes vocales (glotte) qui produisent un son harmonique avec une distribution de l'énergie assez plate en fréquence. Le conduit vocal, les fosses nasales ainsi que la place des articulateurs (langue, mâchoire, lèvres) peuvent être modélisés par un filtre qui modifie le son glottique pour produire le son tel que nous le percevons à la sortie des lèvres d'un locuteur. Ainsi, dans la représentation R_SF, le signal de la parole résulte de la combinaison d'une certaine énergie acoustique (interaction des poumons et du larynx) couplée à une fonction de transfert déterminée par la forme des cavités supra-glottiques. Dans le cadre du traitement du signal, le modèle Source-Filtre décrit le signal de la parole comme la convolution d'un signal d'excitation par un filtre variable dans le temps. L'excitation caractérise la variation de la pression acoustique dans le larynx et le filtre représente le comportement temps-fréquence de la fonction de transfert du conduit vocal.

Pour la plupart, il suffit de modéliser la production d'un signal vocal échantillonné par un modèle de système à temps discret comme le montre la figure 1.3 [30].

Dans ce modèle, l'excitation non voisée est supposée être une séquence de bruit aléatoire, et l'excitation voisée est supposée être un train d'impulsions périodiques avec des impulsions espacées par la période du pitch. Ce système peut être décrit par l'expression de convolution suivante

$$s(n) = \sum_{m=0}^{\infty} h(m)u(n-m) \tag{1.1}$$

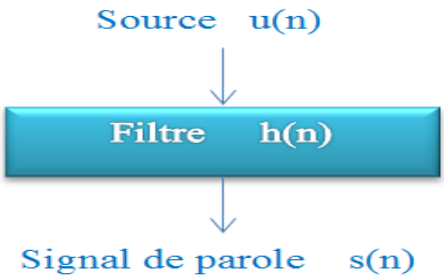


FIGURE 1.3 – Modèle Source-Filtre de la production de la parole. ( s(n) est le signal vocal entier, h(n) est le filtre, et u(n) est la source).

Pour simplifier l'analyse, on suppose souvent que le système est un filtre tout-pôle avec une fonction de transfert du système de la forme

$$H(z) = \frac{G}{(1 - \sum_{k=1}^{p} a_k z^{-k})} \tag{1.2}$$

où $G$ et $p$ sont respectivement le gain et l'ordre du filtre.

Le système linéaire est supposé modéliser les effets du spectre composite du rayonnement, du tube du conduit vocal et de la forme de l'impulsion d'excitation glottale (pour le signal voisé uniquement). Sur un court intervalle de temps le système linéaire dans le modèle est communément appelé simplement le «système du conduit vocal» et la réponse correspondante est appelée "réponse impulsionnelle du conduit vocal".

Pour les systèmes linéaires tout-pôle, comme représenté par l'équation 1.2, l'entrée et la sortie sont liées par une équation de différence de la forme

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{1.3}$$

Cette représentation R_SF est intimement liée au modèle prédictif linéaire qui sera discuté dans la section suivante.

### 1.5.1.2 Représentation predictive linéaire (R_PL)

En 1960, Fant a introduit un modèle linéaire de la forme d'onde du signal vocal dans le domaine temporel [9]. À l'hypothèse source- filtre, il a ajouté l'hypothèse de l'indépendance entre la forme d'onde glottale et le conduit vocal. Le conduit vocal est modélisé comme un filtre tous pôles, également appelé "Autorégressif"(AR).

La figure 1.4 [8] illustre un modèle simplifié de type AR de la production de la parole.

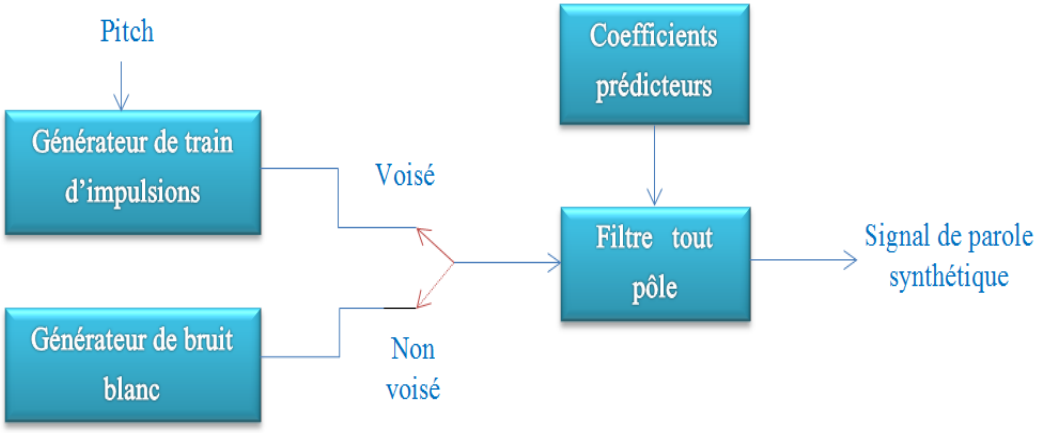


FIGURE 1.4 – Modèle linéaire prédictif de la parole

La forme d'onde glottale est donc modélisée à peu près comme un train d'impulsions avec une période fondamentale égale à la hauteur (pitch) (pour les sons voisés) et comme un bruit blanc avec une moyenne nulle et une variance unitaire (pour les sons non voisés) [30].

Le terme "prédiction linéaire" se réfère au mécanisme d'utiliser une combinaison linéaire des échantillons passés (précédents), $s(n-1)$, $s(n-2)$, ..., $s(n-p)$, pour approximer ou prédire l'échantillon actuel $s(n)$ [8]. Ainsi, un prédicteur linéaire d'ordre $p$, avec des coefficients de prédiction $\alpha_k$, est défini comme un système dont la sortie est

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1.4}$$

Où $\hat{s}(n)$ est le signal prédit.

L'erreur de prédiction est donnée par

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1.5}$$

L'équation 1.5 peut être présentée dans le domaine $z$ comme

$$E(z) = S(z) \times A(z) \tag{1.6}$$

où $E(z)$ est la *Tz* de $e(n)$, $S(z)$ est la *Tz* de $s(n)$, et $A(z)$ est la *Tz* du filtre de l'erreur de prédiction donnée par

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{1.7}$$

Après une inspection plus poussée des équations 1.3 et 1.5, on peut voir que si le modèle est exactement précis pour le signal de parole, et si $\{a_k\}$ = $\{\alpha_k\}$, alors $e(n)$ = $Gu(n)$.

Ainsi, $A(z)$ devient le filtre inverse du système $H(z)$ de l'équation 1.2

$$H(z) = \frac{G}{A(z)} \tag{1.8}$$

## Analyse-synthèse prédictive

Le principal problème de l'analyse prédictive linéaire devient ainsi l'estimation des coefficients prédicteurs $\{a_k\}$ de sorte que l'erreur de prédiction $e(n)$ soit minimisée sous certains critères. L'erreur quadratique moyenne est le critère d'optimisation le plus utilisé. Les coefficients $\{\alpha_k\}$ qui minimisent l'erreur quadratique moyenne sont supposés être les paramètres de la fonction système $H(z)$ de l'équation 1.2.

L'erreur de prédiction au carré $E_n$ dans une trame de courte durée $s_n(m)$ et commençant à l'échantillon $n$ est définie comme

$$E_n = \sum_{m} e_n^2(m) \tag{1.9}$$

$$E_n = \sum_m (s(m) - \hat{s}(m))^2 \tag{1.10}$$

Deux approches majeures du calcul des coefficients de prédiction ont été développées : la méthode d'autocorrélation et la méthode de covariance [41, 42] selon la plage de variation de la variable $m$ dans la relation 1.10. Par conséquent, la minimisation de l'équation 1.10 conduit à des équations normales qui peuvent être résolues en utilisant plusieurs algorithmes.

Pour limiter la longueur du signal à analyser, l'approche d'autocorrélation utilise une fenêtre de pondération ( fenêtre de Hamming ) et la plage de la variable $m$ est ensuite définie sur l'intervalle $(-\infty, \infty)$. La minimisation de l'equation 1.10 mène aux équations de Yule-Walker qu'on peut écrire sous la notation matricielle

$$R.a = r \tag{1.11}$$

où $a = \{a_1, ..., a_p\}^T$ est le vecteur des coefficients cherchés du filtre $A(z)$, $r = \{\hat{r}[1], ..., \hat{r}[p]\}^T$ et $R$ est une matrice de Toeplitz symétrique composée des termes $\hat{r}[k]$. Les termes $\hat{r}[k]$ sont des estimations des premiers $p + 1$ coefficients d'autocorrélation du signal $s[n]$ .

L'avantage principal de l'approche d'autocorrélation est la stabilité du filtre $H(z)$. Son inconvénient est que l'estimation des paramètres du modèle est influencée par l'application de la fenêtre de pondération.

Dans l'approche de covariance, il n'est pas nécessaire d'utiliser une fenêtre de pondération car l'erreur quadratique moyenne est minimisée sur un intervalle de longueur finie. Par contre la stabilité du filtre $AR$ n'est pas garantie par cette méthode et elle est un peu plus complexe que la méthode d'autocorrélation. C'est pour cette raison que l'approche de covariance est moins utilisée que la méthode d'autocorrélation. L'avantage majeur de la méthode de covariance est qu'elle fournisse des paramètres du filtre modélisant l'enveloppe spectrale avec une précision légèrement meilleure que ceux fournis par la méthode d'autocorrélation .

L'analyse prédictive linéaire est appliquée trame par trame au signal de la parole. Ainsi, pour chaque trame, un filtre prédictif linéaire est généré. Ce filtre modélise la forme de l'impulsion d'ex-

citation glottale, les effets du conduit vocal et des radiations des lèvres.

Enfin, la synthèse prédictive linéaire est réalisée comme suit : lors de la parole voisée, un simple train d'impulsions excite le filtre prédictif linéaire, et pour le signal de la parole non voisée, le filtre est excité par un bruit blanc.

La méthode d'analyse prédictive linéaire a été l'une des techniques d'analyse de la parole les plus puissantes car elle est simple, rapide et possède un nombre limité de paramètres et elle a été appliquée avec succès dans l'analyse-synthèse de la parole [11]. Le principal inconvénient de cette méthode est qu'elle est intrinsèquement buzzy en raison de sa nature paramétrique, et cela dégrade la qualité de la parole. De même, les phonèmes tels que les nasales ne peuvent pas être modélisés par le modèle de prédiction linéaire car ils contiennent des antiformants, et ce modèle est un modèle tout-pôle. Afin d'améliorer la qualité de la synthèse de prédiction linéaire certains efforts ont été consacrés en adoptant un modèle d'excitation complexe plus approprié. Par conséquent, des variantes du modèle de prédiction linéaire de base ont été développées tel que "Multipulse Linear Predictive Coding" [43].

**Exemple de la représentation spectrale predictive linéaire du signal de la parole**

Comme illustré à la figure 1.5 et afin de démontrer la performance la représentation R_PL pour la modélisation de la parole, nous avons calculé les spectres de puissance, à partir des échantillons du signal vocal par une transformée de Fourier, et à partir de la fonction de transfert d'un modèle linéaire prédictif, et ce pour un segment de la parole voisée.

## 1.5.2 Représentation fréquentielle du signal de la parole

### 1.5.2.1 Représentation par la transformée de Fourier à court terme (TFCT)

La TFCT [44, 45] est une technique non paramétrique utilisée pour définir les caractéristiques fréquentielles du signal de la parole habituellement représenté dans le domaine temporel. Le modèle sinusoïdale est considéré comme une extension paramétrique de la TFCT [46].

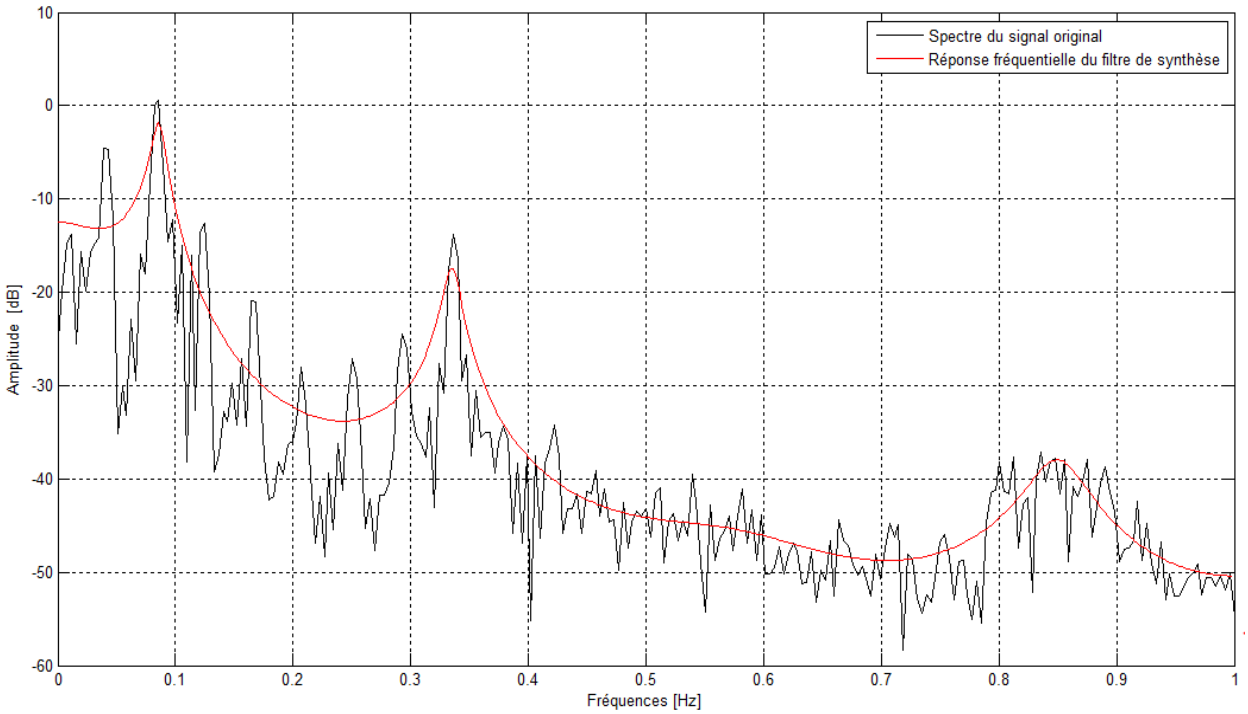


FIGURE 1.5 – Représentation spectrale predictive linéaire du signal de la parole.

Les représentations utilisant la TFCT sont connues sous le nom de représentations spectrales et elles ont été beaucoup utilisées dans les applications de traitement du son car elles imitent le comportement de l'oreille humaine [47, 48]. On peut considérer donc que le système auditif comme un analyseur de spectre, détectant les fréquences présentes dans le son entrant à chaque instant. Le spectre obtenu par le système auditif est à l'échelle logarithmique alors que l'analyse traditionnelle de Fourier calcule le spectre avec une échelle linéaire. Il a été démontré que la TFCT [44, 45] est une technique très générale utile dans l'étude des signaux variant dans le temps, tels que les sons de la parole.

La transformé de Fourier (TF) consiste en la décomposition d'une forme d'onde en un nombre de composantes sinusoïdales comme suit :

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \exp(-j\omega t) dt \tag{1.12}$$

où $t$ est l'indice de temps continu en secondes et $\omega$ est l'indice de fréquence continu exprimé en radians par seconde. Étant donnée une forme d'onde continue $x(t)$, la transformée de Fourier renvoie $X(\omega)$. Il est usuel d'interpréter $X(\omega)$ Comme le spectre de fréquence entier.

Il suffit de dire que $X(\omega)$ est une fonction périodique de $\omega$ avec la période $2\pi$ et qu'il est possible de récupérer la fonction originale $x(t)$ au moyen de la TF inverse,

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) \exp(j\omega t) d\omega \tag{1.13}$$

Autrement dit, étant donné le spectre $X(\omega)$, La TF inverse renvoie la forme d'onde $x(t)$ à partir de laquelle elle a été obtenue. Ainsi, la TF est considérée comme un système identité.

Dans le domaine discret, la TF continue devient la TF discrète (TFD). Si $x(n)$ est un signal de longueur $N$, alors sa TFD est

$$X(k) = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} x(n) \exp(-j\omega_k n), \omega_k = \frac{2\pi k}{N}, Npair, k = 0, 1, ..., N-1 \tag{1.14}$$

Où $\omega_k$ est la fréquence discrète en radian/seconde, $n$ est l'indice de temps discret, et $k$ l'indice de fréquence discrète. $X(k)$ est le spectre discret. La relation entre la fréquence en radian et la fréquence en Hertz (Hz) est donnée par

$$f = f_s \frac{\omega}{2\pi} \tag{1.15}$$

où $f$ la fréquence en Hz, $f_s$ est la fréquence d'échantillonnage et $\omega$ la fréquence en radian/seconde.

Puisque l'indice de fréquence $k$ est discret, la TFD suppose que $x(n)$ peut être représenté par un nombre fini de sinusoïdes, donc le signal $x(n)$ est limité en fréquence. Les fréquences des sinusoïdes sont équidistantes entre 0 Hz et la fréquence d'échantillonnage $f_s$, ou en radians entre $0$ et $2\pi$. Ainsi, la TFD prend une séquence de longueur $N$ (le signal de temps) et produit une autre séquence de longueur $N$ (le spectre de fréquence). La TDF inverse est alors

$$x(n) = \frac{1}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} X(k) \exp(j\omega_k n) \tag{1.16}$$

Autrement dit, étant donné le spectre discret $X(k)$, la TFD-inverse renvoie la forme d'onde discrète $x(n)$ à partir de laquelle il a été obtenu.

La prévalence de la TFD dans tant d'applications différentes est due à l'existence d'un algorithme très rapide pour son calcul appelé FFT (Fast Fourier Transform). L'implementation traditionnelle de la FFT nécessite que la longueur du signal, N, soit une puissance de 2. En faisant cette

restriction, le temps de calcul est alors réduit.

Le modèle correspondant à la TFCT peut être formulé comme suit :

$$s(n) = \frac{1}{N}\sum_{l=0}^{L-1}\int_{-\pi}^{\pi} X_l(\omega)\exp(j\omega_k(n-lH))d\omega \tag{1.17}$$

$X_l(\omega)$ est le spectre continu d'un signal $x(n)$ à la trame $l$. C'est-à-dire que le signal $s(n)$ est modélisé comme la somme d'une série de TF-inverses.

En pratique, les sons de la parole sont des formes d'onde non périodiques et variables dans le temps, caractéristiques pour lesquelles la TF n'est pas appropriée. L'alternative est d'utiliser la TFCT. Dans le cas discret, elle peut être définie comme

$$X_l(k) = \sum_{n=0}^{N-1} w(n)x(n+lH)\exp(-j\omega_k n) \qquad l = 0, 1, ... \tag{1.18}$$

où $w(n)$ est une "fenêtre" réelle qui détermine la portion $l$ du signal d'entrée $x(n)$. $H$ est le taux d'avancement de la fenêtre.

La sortie de la TFCT est une série de spectres, un pour chaque trame $l$ de la forme d'onde d'entrée. Chaque spectre $X_l(k)$ est une fonction de valeur complexe. Une représentation plus utile est en termes de magnitudes et de phases, obtenues par

$$A_k = |X_l(k)| = \sqrt{a^2(k) + b^2(k)} \tag{1.19}$$

$$\Theta_l(k) = \angle X_l(k) = tan^{-1}(\frac{b(k)}{a(k)}) \tag{1.20}$$

où $|X_l(k)|$ est la magnitude, $\Theta_l(k)$ est la phase, et ($a(k)$, $b(k)$) sont les parties réelles et imaginaires de la valeur complexe renvoyée par la TF,

$$a(k) = \Re\{X_l(k)\}, b(k) = \Im\{X_l(k)\} \tag{1.21}$$

Il y a plusieurs problèmes dans le calcul de la TFCT qui, nécessitent une attention particulière : par exemple, le choix de la fenêtre d'analyse, le calcul de la TFD, et la taille du temps d'avancement de la fenêtre.

Les représentations "phase-vocoder" [28, 49] étaient les versions les plus connues de la TFCT utilisées pour le traitement du son.

**Exemple d'un spectrogramme d'un segment de la parole**

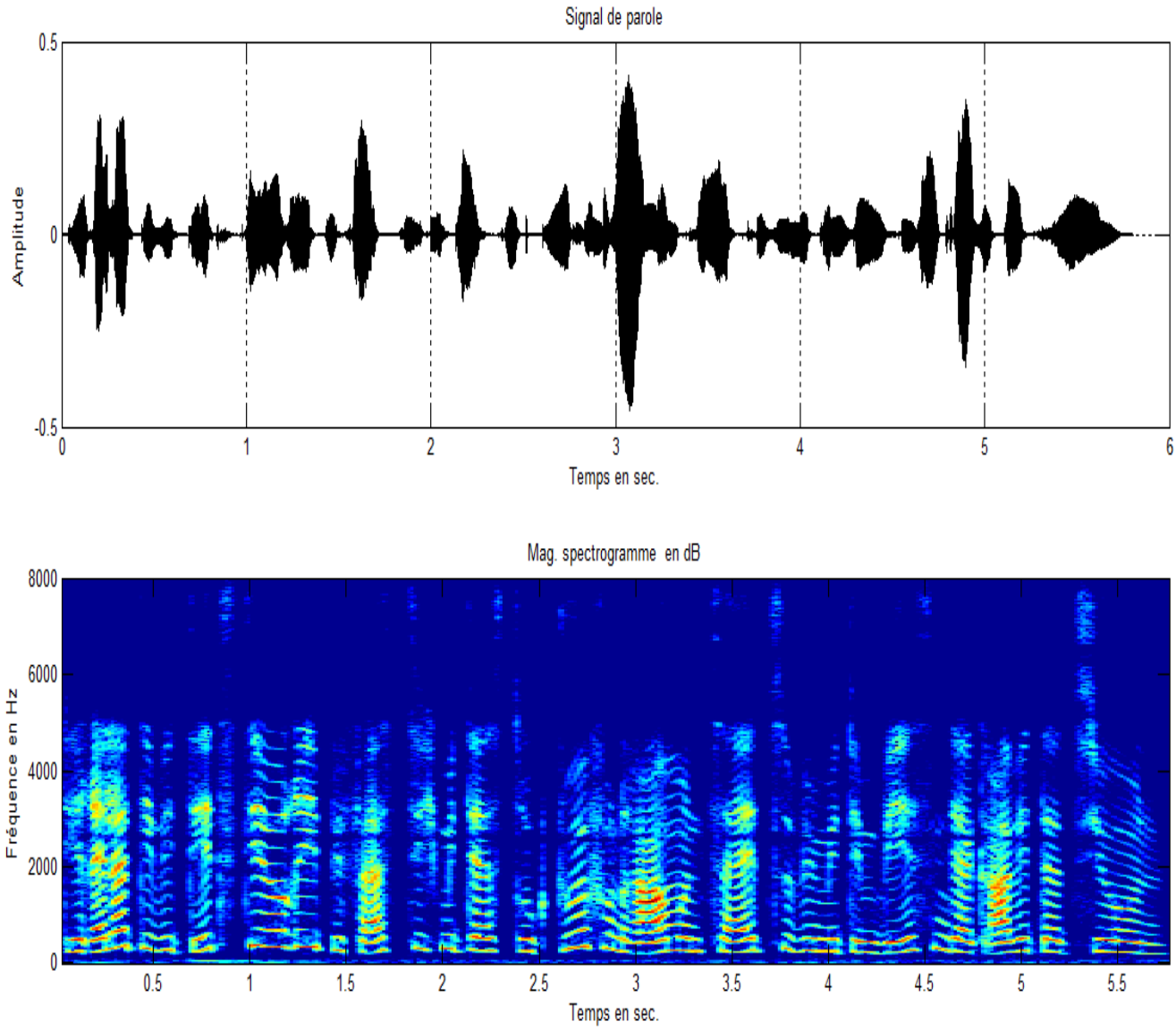


FIGURE 1.6 – Spectrogramme d'un segment de la parole

Pour mettre en évidence les caractéristiques du signal de la parole, l'étude de l'évolution temporelle et fréquentielle du signal est requise, en utilisant les méthodes modernes de traitement du signal. Par exemple le spectre de puissance à court terme du signal vocal est obtenue grâce la technique de la TF et l'évolution temporelle de ce spectre aboutit au spectrogramme de ce signal. La TFCT d'un signal est donc un outil principal du traitement du signal qui permet de caractériser les propriétés fréquentielles des segments phonétiques de la parole.

Pour représenter un signal vocal dans le domaine spectrale, on commence par segmenter le

signal de parole pour obtenir une trame composée de N échantillons puis on utilise une fenêtre de pondération (Hamming, Hanning, ...) pour atténuer les distortions spectrales introduites par l'extraction de la trame du signal de parole. On calcule enfin la TFD des échantillons de la trame pondérée.

Le spectrogramme de la figure 1.6 [50]met bien en évidence l'évolution temporelle des structures fréquentielles, formantiques des parties voisés d'un segment de parole.

# Conclusion

Dans ce chapitre, nous avons présenté un bref aperçu du fonctionnement d'un système de synthèse de la parole à partir du texte (TTS) ainsi que les plus importantes techniques de synthèse utilisées par ce dernier, à savoir, la synthèse par formants, la synthèse articulatoire, la synthèse concaténative, la synthèse par sélection d'unité et enfin la synthèse par modèle de Markov caché. Nous avons donné un bref historique des représentations les plus populaires du signal de la parole. Nous avons aussi passé en revue le mécanisme de la production de la parole ainsi que les principales représentations utilisées par les techniques de synthèse pour bien modéliser le signal de la parole, à savoir, la représentation source-filtre, la représentation prédictive linéaire et la représentation par transformée de Fourier. Nous avons également donné des exemples illustratifs utilisant ces différents types de représentations.
Dans le chapitre suivant, nous présenterons un autre type de représentation du signal de la parole nommée la représentation sinusoïdale stationnaire et ses variantes.

# Chapitre 2

# Représentations sinusoïdales stationnaires du signal de la parole

La méthode de prédiction linéaire a été la représentation prédominante pour estimer les paramètres de base de la parole (par exemple : la hauteur, les formants, le spectre) et pour représenter la parole pour une transmission ou un stockage à faible débit binaire, jusqu'à la fin des années quatre vingt. Depuis ce temps, des représentations plus complexes qui offraient une meilleure qualité du signal, tel que la représentation sinusoïdale et ses dérivées sont apparues.

Nous allons décrire dans ce chapitre plusieurs types de représentations sinusoïdales utilisées par les systèmes d'analyse-synthèse du son ou du signal de la parole. Dans la littérature, il existe deux approches différentes utilisées par les systèmes d'analyse-synthèse : l'approche uniforme et l'approche hybride. Dans l'approche uniforme, les systèmes d'analyse-synthèse traitent toutes les parties du signal vocal de la même manière, en tant que somme de sinusoïdes variant dans le temps. Par contre dans l'approche hybride, les systèmes d'analyse-synthèse décomposent la parole en deux composantes, généralement nommées partie déterministe et partie stochastique. Dans ce chapitre, nous présentons une brève description des plus importantes représentations sinusoïdales stationnaires du signal de la parole utilisées par ces deux types d'approches (uniformes et hybrides) ainsi que leurs processus d'analyse-synthèse. Ce chapitre se termine par une présentation de quelques exemples expérimentaux d'analyse-synthèse de signaux de paroles voisées et une conclusion.

## 2.1 Représentations sinusoïdales uniformes

### 2.1.1 Représentation sinusoïdale (R_S)

Dans la représentation sinusoïdale connue sous le nom "Sinusoidal Model (SM)" [31, 32, 33, 10] , le modèle d'excitation binaire voisée / non voisée de la représentation $R\_PL$ décrite dans le chapitre précédent est remplacé par une somme des fonctions sinusoïdales évoluant dans le temps. Ainsi, le signal de parole est toujours supposé être la sortie d'un filtre numérique variant lentement dans le temps avec une excitation qui capture la nature de la distinction voisée / non voisée dans la production de la parole (Excitation exprimée comme une somme de sinusoïdes).

La représentation R_S est dès l'origine utilisée pour la génération et la transformation des sons (parole ou music). On peut dire que la représentation R_S est une application du théorème de Fourier qui montre que tout signal périodique peut être modélisé par une somme de sinusoïdes avec différentes fréquences et amplitudes.

Pour l'analyse sinusoïdale, de nombreuses approches ont été proposées dans la littérature pour l'estimation des paramètres des représentations sinusoïdales. On peut citer deux importantes techniques d'estimations : La première utilise une analyse basée sur la transformée de Fourier (TF) et la deuxième est basée sur l'analyse des moindres carrés (MC)[4].

Pour la synthèse sinusoïdale, deux importantes techniques existent. La première catégorie utilise une technique d'interpolation de paramètres du modèle entre les trames successives avant la synthèse proprement dite et qui consiste à sommer toutes les composantes sinusoïdales (la synthèse par interpolation). La deuxième technique utilise le principe de chevauchement et addition de sinusoïdes (Synthèse type overlap-add) [4].

En essayant de représenter la parole par des modèles sinusoïdaux, plusieurs approches ont été proposées [31, 32, 33, 10]. La plus populaire et la plus célèbre représentation a été présentée dans [10]. Cette représentation s'est avérée plus générale que les représentations sinusoïdales antérieures.

Le modèle $SM$ proposé dans [10] est basé sur la $TFCT$ et caractérisé par les amplitudes, les fréquences et les phases des pics les plus importants dans le temps dans la série de spectres renvoyés par la $TFCT$. A partir de cette représentation, un son est généré en synthétisant une onde sinusoïdale pour chaque trajectoire de pic trouvé. Ainsi, nous pouvons interpréter la représentation R_S comme une simplification de la sortie de la $TFCT$, où seuls les pics spectraux pertinents sont pris en compte dans l'ensemble des spectres renvoyés par la $TFCT$. Ces pics, qui représentent chacun une sinusoïde, sont ensuite regroupés en trajectoires fréquentielles.

Le signal de la parole résultant du modèle complet est écrit comme suit

$$s(t) = \sum_{l=1}^{L} a_l(t) \cos(\phi_l(t)) \tag{2.1}$$

avec

$$\phi_l(t) = \int_0^t \omega_l(\tau) d\tau \tag{2.2}$$

où $L$ est le nombre des sinusoïdes ; $a(t)$, $\phi(t)$ représentent respectivement amplitude et la phase de la sinusoïde. $\omega(t)$ est la fréquence instantanée en radian.

La figure 2.1 [10] montre un schéma général d'un système d'analyse-synthèse basé sur le modèle $SM$.

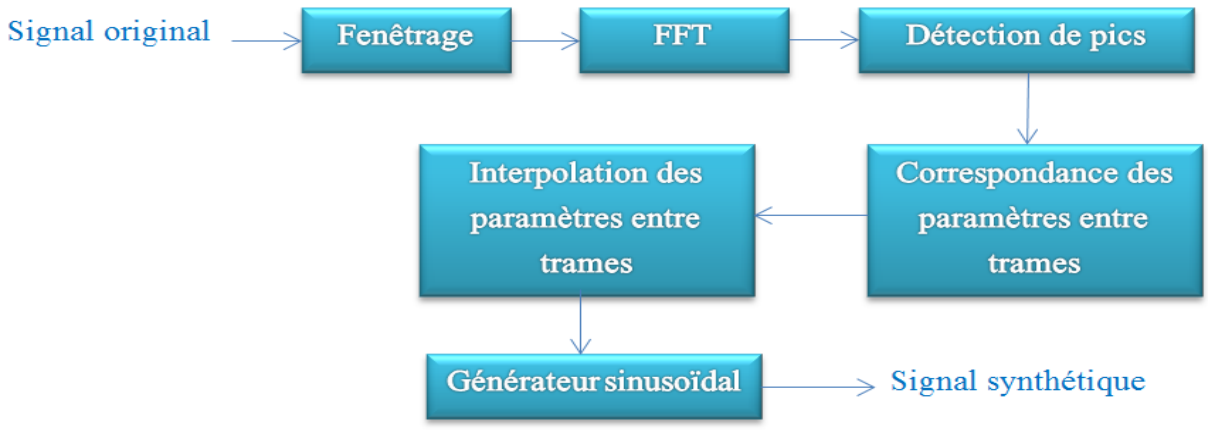


FIGURE 2.1 – Bloque diagramme simplifié du modèle sinusoïdal

En bref et d'après la figure 2.1, le système d'analyse-synthèse sinusoïdale commence par calculer la $TFCT$. Ensuite, à partir des spectres de magnitude et de phase renvoyés par la $TFCT$, une série de trajectoires de pics sont extraites par un algorithme de détection de pics. Chaque trajectoire de pic représente donc une sinusoïde caractérisée par des paramètres variables dans le temps (amplitudes, fréquences, phases). Ensuite un algorithme de poursuite de pics est utilisé par le système. La partie de synthèse du système utilise les trajectoires de pics pour générer des ondes sinusoïdales qui sont additionnées pour créer la forme d'onde synthétisée finale.

## Analyse-synthèse sinusoïdale

Donnons maintenant plus de détails sur les étapes d'analyse-synthèse sinusoïdale. A l'étape de l'analyse, il est nécessaire d'estimer le nombre de composantes sinusoïdales, leurs amplitudes et leurs fréquences. À cette fin, la TFCT est utilisée. Ensuite, pour chaque trame, les pics spectraux sont obtenus en recherchant tous les maxima locaux sur le spectre d'amplitude en éliminant ceux dont l'amplitude est inférieure à un seuil donné. Il est important d'avoir les pics aussi bien résolus que possible. Il a été démontré qu'un « zéro-padding » donne un spectre plus lisse, ce qui rend la détection des pics plus facile et plus précise. Ici, le facteur de « zéro-padding » devrait être aussi grand que possible. La position des pics fournit les fréquences et les amplitudes des composants sinusoïdaux. Les phases de ces composantes sont calculées comme la phase de la transformée de Fourier à court terme pour une fréquence donnée. Pour chaque trame, on obtient ainsi un ensemble de L pics spectraux.

Étant donné que le nombre de pics spectraux n'est pas constant (avec des amplitudes et des fréquences changeant lentement) l'étape suivante consiste à utiliser un algorithme de continuation ou de poursuite de pics qui a pour tâche l'assignation des pics aux trajectoires de fréquences en faisant correspondre les pics de la trame précédente avec celle en cours. Ces trajectoires sont « nées-born," », ou « tuées- killed" » à n'importe quelle trame en augmentant l'amplitude de / ou vers 0. Dans le cas où une composante sinusoïdale est née (born) ou en train de mourir (dying), l'amplitude instantanée s'annule linéairement jusqu'à l'instant d'analyse suivant tandis que la fréquence instantanée reste constante jusqu'à la disparition de la composante.

L'algorithme de poursuite des pics renvoie les valeurs des pics prédominants organisés dans les trajectoires de fréquence. Chaque pic est une triade ($\hat{a}_l^k$, $\hat{\omega}_l^k$, $\hat{\phi}_l^k$) où $k$ est le numéro de la trame et $l$ est le numéro de piste auquel il appartient.

Le processus de synthèse prend ces trajectoires, et calcule une trame du son synthétisé $s^k(n)$ en utilisant

$$s^k(m) = \sum_{l=1}^{L^k} \hat{a}_l^k(t) \cos(m\hat{\omega}_l^k + \hat{\varphi}_l^k) \tag{2.3}$$

où $L^k$ est le nombre de trajectoires présentes à la trame $k$

Le son final $s(n)$ résulte de la juxtaposition de toutes les trames de synthèse (c.à.d, qu'il n'y a pas de chevauchement).

Pour éviter les "cliques" aux limites de la trame, les paramètres ($\hat{a}_l^k$, $\hat{\omega}_l^k$, $\hat{\phi}_l^k$) sont interpolés d'une trame à l'autre. L'amplitude instantanée $\hat{a}(m)$ est facilement obtenue par interpolation linéaire. Les valeurs de fréquence et de phase sont liées (la fréquence est la dérivée de la phase) et elles sont interpolées en utilisant une fonctions cubique [10].

Donc, pour obtenir le signal synthétique final, on doit générer une onde sinusoïdale pour chaque trajectoire de fréquence, et les additionner toutes. L'amplitude instantanée et la phase pour chaque onde sinusoïdale sont calculées en interpolant les valeurs d'une trame à l'autre.

Il a été montré que la représentation R_S fournit une reconstruction très précise de la parole voisée et a été appliquée avec succès dans la synthèse de parole [51, 52] .

### 2.1.2 Représentation ABS/OLA

La recherche présentée dans [53, 54] a étudié la possibilité d'utilisation d'une procédure d'analyse par synthèse (ABS) pour déterminer les paramètres d'une formulation d'un modèle sinusoïdal à chevauchement et addition (OLA).

Le modèle proposé pour représenter $s[n]$ est donc une formulation de modèle sinusoïdal $OLA$

donné sous sa forme la plus générale par

$$\hat{s}[n] = \sigma[n] \sum_{k=-\infty}^{+\infty} w_s[n - kN_s]\hat{s}^k[n - kN_s] \tag{2.4}$$

La fenêtre de synthèse $w_s(n)$ est une fenêtre complémentaire obéissant à la contrainte

$$\sum_{k=-\infty}^{+\infty} w_s[n - kN_s] = 1 \tag{2.5}$$

Pour tout $n$, $N_s$ détermine la longueur de la trame de synthèse.

La contribution synthétique $\hat{s}^k[n]$, est donnée par

$$\hat{s}^k[n] = \sum_{j=1}^{J[k]} A_j^k \cos(2\pi f_j^k n/F_s + \phi_j^k) \tag{2.6}$$

où $0 < f_j^k < F_s/2$, et la séquence d'enveloppe $\sigma[n]$ reflète les variations de l'énergie de $s[n]$ dans le modèle, afin d'augmenter la précision dans les régions transitoires de $s[n]$. Donc, $\hat{s}[n]$ est une somme de formes d'onde synthétiques pondérées par des fenêtres chevauchées par $N_s$ échantillons, additionnées et modulées par $\sigma[n]$, où chaque forme d'onde synthétique est produite en additionnant des sinusoïdes d'amplitudes, de fréquences et de phases différentes .

## Analyse-synthèse ABS/OLA

Comme avec toute approche basée sur la modélisation de la parole, il faut prendre soin de choisir $N_s$ de telle sorte que le signal vocal puisse être supposé stationnaire sur un intervalle de trame donnée. Les valeurs typiques correspondent à des valeurs entre 5 et 20 msec, selon les exigences de l'application.

L'ensemble de paramètres qui doit être déterminé pour représenter $s[n]$ est constitué de la séquence d'enveloppe $\sigma[n]$ et des amplitudes $A_j^k$, des fréquences $\omega_j^k$ et des phases $\phi_j^k$ de chaque séquence de contribution synthétique $\hat{s}^k[n]$.

La détermination d'une envelope $\sigma[n]$ est la premiere étape à effectuer. Étant donné $\sigma[n]$, l'objectif de l'analyse est de déterminer les paramètres d'amplitude, de fréquence et de phase pour

chaque $\hat{s}^k[n]$ dans l'équation 2.4 tel que $\hat{s}[n]$ soit « le plus proche » de $s[n]$ dans un certain sens. Une approche typiquement employée pour résoudre des problèmes de ce type consiste à minimiser l'erreur quadratique moyenne suivante

$$E = \sum_{n=-\infty}^{\infty} (s[n] - \hat{s}[n])^2 \tag{2.7}$$

en termes de paramètres de $\hat{s}[n]$. Cependant, tenter de résoudre ce problème simultanément pour tous les paramètres n'est pas pratique.

Heureusement, si $s[n]$ est approximativement stationnaire sur de courts intervalles de temps, il est possible de résoudre le problème pour les paramètres d'amplitude, de fréquence et de phase de $\hat{s}^k[n]$ isolément en approximant $s[n]$ sur une trame d'analyse de longueur $2N_a + 1$ échantillons centré à $n = kN_s$. La contribution synthétique $\hat{s}^k[n]$ peut alors être déterminée en minimisant

$$E^k = \sum_{n=-N_a}^{N_a} w_a[n](s[n + kN_a] - \sigma[s[n + kN_a]\hat{s}^k[n])^2 \tag{2.8}$$

en termes des amplitudes, des fréquences et des phases de $\hat{s}^k[n]$.

Le but donc de l'analyse-par-synthèse (ABS) est de mettre à jour l'approximation de $s[n]$ en ajoutant un seul composant telle que l'approximation de mise à jour soit aussi bonne que possible.

Il a été montré dans [53, 54] que la représentation $ABS/OLA$ fournie une bonne qualité de synthèse vocale comparée avec la représentation $R\_S$ suggérée dans [10].

## 2.2 Représentations sinusoïdales hybrides

Les représentations sinusoïdales décrites ci-dessus sont très adaptées à la modélisation des signaux de paroles périodiques. En effet dans ce cas particulier, un faible nombre de sinusoïdes est requis pour représenter ce type de signaux. Par contre pour représenter les signaux de la parole bruités, ces représentations, bien que toujours applicable, deviennent beaucoup moins adaptées, car un grand nombre de composantes sinusoïdales est alors requis.

Dans la littérature, la séparation des composantes périodiques et apériodiques de la parole a gagnée beaucoup d'intérêt pour la recherche car les représentations sinusoïdales présentées dans les

sections précédentes ne sont pas appropriées pour la manipulation de sons contenant des composants de bruit. Plusieurs autres techniques ont été proposées au cours des dernières décennies, afin de fournir des représentations plus flexibles et de haute qualité via une combinaison de sinusoïdes et de bruit.

Un organigramme typique d'un système hybride est illustré à la figure 2.2

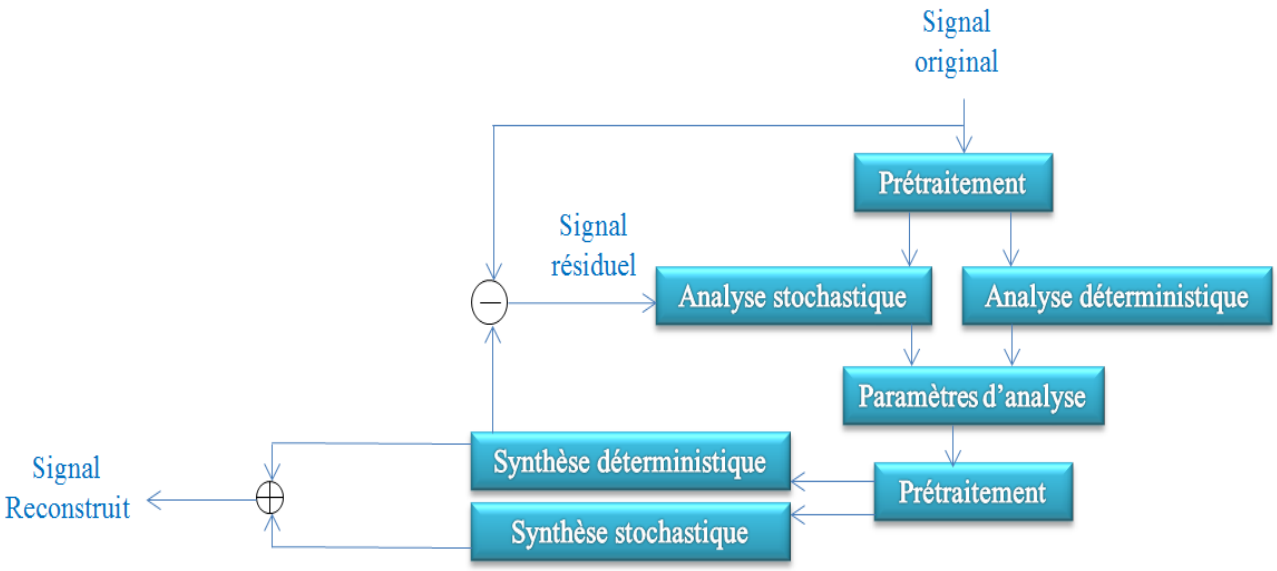


FIGURE 2.2 – Bloque diagramme simplifié de la représentation hybride

Discutons brièvement les éléments d'un système hybride général. Tout d'abord, dans la partie analyse, l'étape de prétraitement inclut souvent des actions telles que l'estimation de la fréquence fondamentale, la décision voisée / non voisée, l'estimation de la fréquence de voisement maximale ($Fm$), filtrage, amélioration ou annulation de bruit. La partie déterministe est responsable de la modélisation des caractéristiques déterministes de la parole, tandis que la partie stochastique modélise la composante aléatoire du signal de la parole, comme le bruit de frottement, la parole non voisée, etc.

Lorsque les paramètres d'analyse pour tous les composants vocaux sont estimés, ils sont transmis à l'étape de synthèse, où un prétraitement des paramètres est effectué, comme par exemple l'interpolation de paramètres ou l'estimation d'enveloppe spectrale, en cas de modifications de la parole. Enfin, chaque composant est synthétisé séparément et tous les composants sont additionnés

pour former le signal vocal synthétisé.

Les systèmes hybrides sont considérés comme bien adaptés à la resynthèse et aux modifications prosodiques, puisqu'une séparation bien maîtrisée de la parole en une composante déterministe et une composante stochastique conduit à une meilleure manipulation et améliore la qualité de la synthèse et des modifications de la parole [12].

Des exemples typiques de tels systèmes hybrides sont brièvement décrit dans les sections qui suivent

### 2.2.1 Représentation déterministique plus résiduel (R_DR)

Dans cete section, un modèle alternatif au modèle SM a été introduit dans [55, 56] qui considère qu'un son est composé d'une partie déterministe plus un résidu.

Un signal déterministe est traditionnellement défini comme tout ce qui n'est pas du bruit (c.à.d, une partie parfaitement prévisible, prévisible à partir de mesures sur tout intervalle continu). Cependant, dans [55, 56], la classe des signaux déterministes considérés est limitée aux sommes des composantes quasi sinusoïdales (Sinus avec variation linéaire d'amplitude et de fréquence par morceaux). Chaque sinusoïde modélise une composante quasi sinusoïdale du son original et c'est un élément indépendant qui peut être synthétisé par lui même. La composante déterministe modélise donc les partiels (Un partiel est une composante sinusoïdale d'un son qui correspond généralement à un mode de vibration du système sonore producteur) du son. Le résidu est alors défini comme la différence entre la partie déterministe originale et la partie déterministe estimée. La somme des deux composants donne le son original.

Dans le système proposé dans [55, 56], le modèle R_DR considère une forme d'onde $s(t)$ comme la somme d'une série de sinusoïdes plus un résiduel $e(t)$,

$$s(t) = \sum_{l=1}^{L} a_l(t) \cos(\phi_l(t)) + e(t) \tag{2.9}$$

où $L$ est le nombre de sinusoïdes, $a_l(t)$ est l'amplitude instantanée et $\phi_l(t)$ la phase instantanée. Le résidu est la différence entre le signal original et la partie déterministe. Cette composante détermi-

niste est définie de la même manière que dans le modèle sinusoïdal, donc la phase instantanée $\phi_l(t)$ est définie par l'équation 2.2.

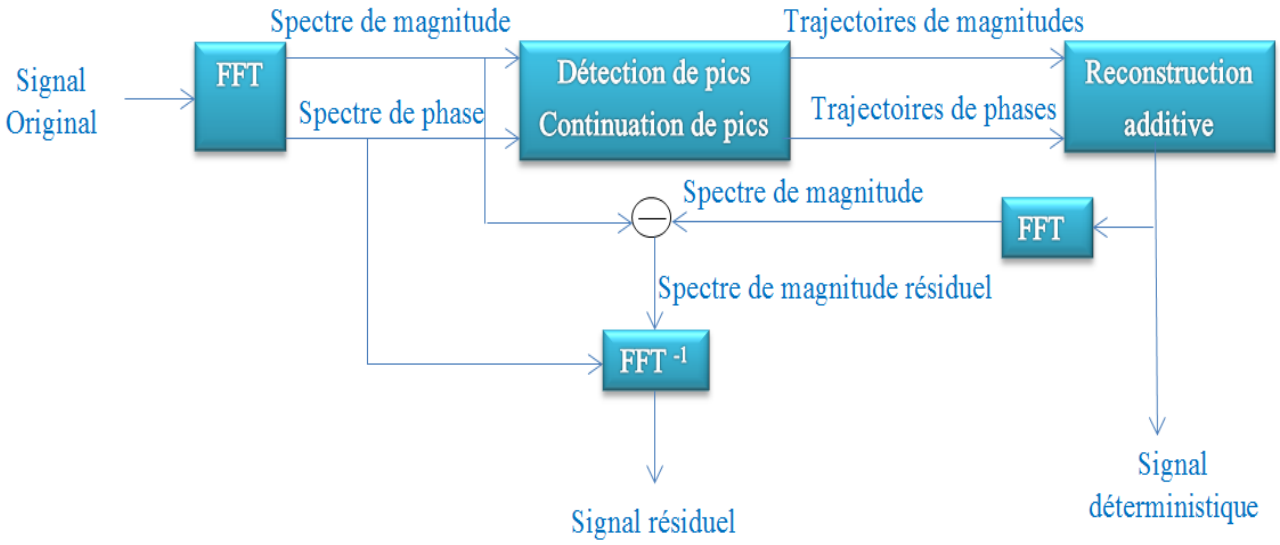


FIGURE 2.3 – Bloque diagramme simplifié du modèle sinusoïdal plus résiduel

Un schéma général de la méthode est montré à la figure 2.3 [55, 56]. Le processus commence en calculant un ensemble de spectres de magnitude et de phase avec la $TFCT$. À partir de ces spectres, on trouve des trajectoires sinusoïdales avec des algorithmes de détection de pics et de continuation de pics. Pour tenir compte du concept du « signal déterministe », dans ce type de représentation le comportement des trajectoires de pics est plus restreint que dans la représentation uniforme R_S, et un nouvel algorithme de continuation de pic a été présenté à cet effet. Les sinusoïdes sont ensuite générées avec une synthèse additive, et le résidu est calculé simplement en soustrayant le signal déterministe de la forme d'onde originale.

Les seules différences entre la représentation hybride R_DR et la représentation uniforme R_S sont donc les suivantes :

— On garde maintenant un signal résiduel, et

— maintenant les sinusoïdes sont limités à être stable (c.à.d, suivre des composants quasi-sinusoïdaux stables), elles modélisent donc uniquement les partiels du son.

La partie déterministe est maintenant plus contraignante, mais le nouveau signal résiduel augmente la généralité du modèle par rapport au modèle sinusoïdal.

### 2.2.2 Représentation déterministe plus stochastique (R_DS)

En allant plus loin dans le modèle déterministe plus résiduel présenté dans la section précédente, une représentation plus flexible et utile pour la manipulation du son a été proposée dans [55, 56].

La représentation résultante comporte deux parties :

1. Une série de fonctions de fréquence et d'amplitude pour la composante déterministe,
2. une série d'enveloppes de spectre d'amplitude pour la partie stochastique du son.

Le signal déterministe est généré à partir des fonctions d'amplitude et de fréquence avec une synthèse additive. Le composant stochastique est créé en effectuant la $TFCT$ inverse de toutes les enveloppes spectrales. La somme des deux formes d'onde résultantes est, pour beaucoup de sons, perceptuellement très proche du signal original.

Comparé au système de la section précédente, il y a un gain dans la flexibilité de la représentation en échange de la propriété d'identité du processus. Avec le modèle déterministe et résiduel, tout son était représenté ; d'autre part, avec le modèle déterministe et stochastique, tous les sons ne peuvent pas être ajustés par le modèle.

Le modèle R_DS est basé sur une modification du modèle R_DR présenté à la section précédente comme suit

$$s(t) = \sum_{l=1}^{L} a_l(t) \cos(\phi_l(t)) + e(t) \tag{2.10}$$

où $L$ est le nombre de sinusoïdes, $a_l(t)$ est l'amplitude instantanée. $e(t)$ est signal résiduel.

La phase instantanée $\phi_l(t)$ est toujours considérée comme l'intégrale de la fréquence instantanée, comme indiquée par l'équation 2.2

La simplification du résidu $e(t)$ est basée sur l'hypothèse qu'il s'agit d'un signal stochastique. Une telle hypothèse permet de modéliser le résidu sous forme de bruit blanc filtré,

$$\hat{e}(t) = \int_0^t h(t, t-\tau) b(\tau) d\tau \tag{2.11}$$

Où $b(t)$ est un bruit blanc et $h(t)$ est la réponse impulsionnelle d'un filtre variant lentement dans le temps (à l'instant $t$, la réponse impulsionnelle est $h(t,)$).

En d'autre termes, le résidu est modélisé par la convolution d'un bruit blanc avec un filtre de mise en forme de fréquence comme suit

$$\hat{e}(t) = h(t) * b(t) \tag{2.12}$$

Le filtrage d'un signal de bruit peut être réalisé en prenant la TF inverse de la réponse en fréquence du filtre multipliée par un terme de phase aléatoire. Cette dernière approche a été utilisée pour synthétiser le signal stochastique.

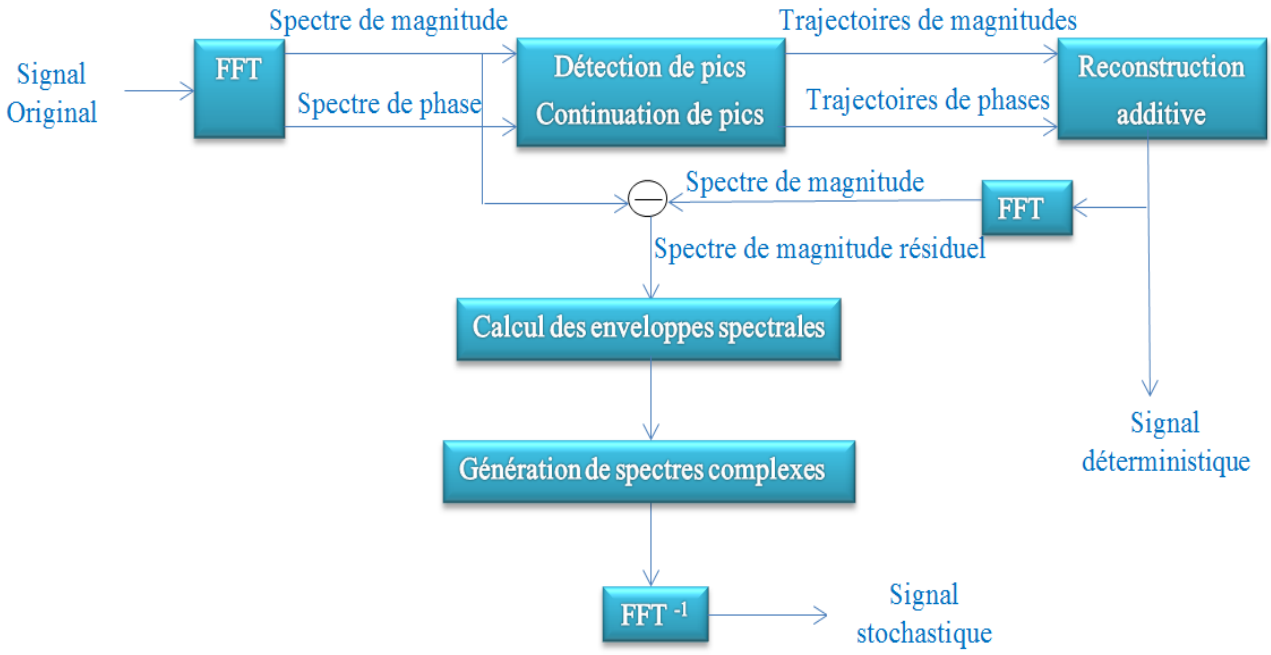


FIGURE 2.4 – Bloque diagramme simplifié du modèle sinusoïdal plus stochastique

La figure 2.4 [55, 56] montre un schéma général d'un système basé sur le modèle R_DS. Premièrement, nous dérivons une série de spectres de magnitude de la forme d'onde en calculant la $TFCT$. Le spectre de phase n'est pas nécessaire et, par conséquent, il n'est pas calculé. Ensuite, nous détectons et suivons les pics prédominants sur chaque spectre, résultant en un ensemble de trajectoires de pic avec une magnitude et une valeur de fréquence pour chaque trame. A partir de ces trajectoires, nous synthétisons la partie déterministe du son en générant une onde sinusoïdale pour chaque trajectoire.

Pour calculer la partie stochastique de la forme d'onde, on obtient d'abord le résidu de spectre d'amplitude. Ceci est fait en calculant le spectre d'amplitude de la composante déterministe et en la soustrayant ensuite du spectre d'amplitude correspondant de la forme d'onde originale. Chaque résidu du spectre d'amplitude est simplifié en y ajustant une enveloppe. L'ensemble résultant d'enveloppes constitue la représentation stochastique. A partir de chaque enveloppe spectrale, le spectre complexe correspondant est généré. Ensuite, la forme d'onde stochastique est synthétisée en effectuant une $TFCT$ inverse en utilisant la méthode « overlap-add ».

### 2.2.3 Représentation Harmonique plus Bruit (R_HB)

Un modèle hybride bien connu pour la parole est le modèle "Harmonic plus Noise (HNM" développé dans [12] et il a été utilisé pour une modification de la parole pour la transformation de la voix. HNM décompose le spectre du signal de la parole en deux bandes : la bande inférieure (partie déterministe) où le signal vocal est modélisé comme une somme de sinusoïdes harmoniquement liées et la bande supérieure (partie stochastique) où le signal vocal est modélisé comme bruit modulé.

L'hypothèse principale derrière la représentation Harmonique plus bruit est donc que le signal vocal est décomposé en une partie harmonique (déterministique) plus une partie de bruit (stochastique). La partie harmonique représente les composantes quasi-périodiques du signal de la parole telles que les voyelles et certaines consonnes voisées et la partie bruit représente la partie non périodique comme le bruit d'aspiration fricative, les explosives, la parole non voisée, etc. La partie harmonique est modélisée à travers un ensemble de sinusoïdes harmoniquement liées avec des amplitudes et des fréquences variant lentement. Cependant, la partie bruit est généralement modélisée

sous la forme d'un bruit blanc Gaussien passant à travers un filtre de mise en forme. Le spectre de la parole est divisé en deux sous-bandes délimitées par une fréquence voisée maximale variable dans le temps.

Par conséquent, le signal de parole dans un tel modèle peut être représenté par l'expression

$$s(t) = s_h(t) + s_n(t) \tag{2.13}$$

Le signal $s_h(t)$ représente la partie harmonique qui est modélisée par :

$$s_h(t) = \sum_{l=1}^{L} A_l \cos(2\pi l f_0 t + \phi_l) \tag{2.14}$$

$$= \sum_{l=1}^{L} a_l \cos(2\pi l f_0 t) + b_l \sin(2\pi l f_0 t) \tag{2.15}$$

où $a_l$, $b_l$, $A_l$, $\phi_l$ sont des réels qui désignent respectivement amplitudes et phase du modèle. $f_0$ est la fréquence fondamentale et $L$ représente le nombre d'harmoniques. Ces paramètres sont supposés constants pour un court intervalle de temps.

Les équations (2.16, 2.17) peuvent être généralisées sous la forme

$$s_h(t) = \sum_{l=1}^{L(t)} A_l(t) \cos(\phi_l(t)) \tag{2.16}$$

avec

$$\phi_l(t) = \int_0^t 2\pi l f_0(\tau) d\tau \tag{2.17}$$

avec $A_l(t)$ , $\phi_l(t)$ et $L(t)$ sont des paramètres variables dans le temps.

## Analyse-synthèse Harmonique plus Bruit

La première étape d'analyse consiste à estimer la partie harmonique, ensuite, la soustraction $s_{hnm}(t) - s_h(t)$ donne un résidu, qui sera d'autant plus proche de la partie bruit. Cette dernière sera elle aussi estimée par une autre technique différente de celle de la partie harmonique. Les deux parties seront ensuite re-synthétisées séparément.

Cependant, le signal $s_n(t)$ qui représente la partie bruitée peut être modélisé comme suit :

$$s_n(t) = e(t)[h(t,\tau) * b(t)] \tag{2.18}$$

où $b(t)$ est un bruit blanc Gaussien ; $h(t)$ est un filtre tout-pôle normalisé variant dans le temps ; $e(t)$ est une fonction enveloppe temporelle d'énergie appliquée pour donner au bruit filtré le motif temporel correct.

Ce modèle est entièrement paramétrique (i.e. : la synthèse se fait exclusivement à partir des paramètres extraits, sans réutiliser d'aucune manière le signal d'origine).

L'estimation du pitch est donc la première étape de l'analyse. A partir de cette estimation initiale du pitch, un modèle harmonique est ajusté (adapté) à chaque trame et la décision voisée / non voisée est faite en utilisant un critère qui prend en compte dans quelle mesure ce modèle harmonique est proche du modèle original. Pour les trames voisées, la fréquence de voisement maximum $f_m$ est alors estimée. Une fois cette fréquence de voisement maximum trouvée, la ré-estimation précise du pitch est nécessaire. Ainsi, Les amplitudes et les phases des harmoniques sont obtenues dans le domaine temporel en utilisant une erreur quadratique pondérée entre la forme d'onde réelle et synthétique, c'est-à-dire, l'estimation des paramètres est effectuée en utilisant la technique des MC,

$$\epsilon = \sum_{t=-N}^{N} (s(t) - s_h(t))^2 \tag{2.19}$$

avec $s(t)$ est le signal original et $s_h(t)$ est le signal harmonique.

Pour l'estimation des paramètres de la composante de bruit (partie non voisée), dans chaque trame d'analyse, une densité spectrale du signal original est modélisée par un filtre $AR$. Ce filtre sera excité par un bruit blanc et les caractéristiques dynamiques sont considérées en utilisant une enveloppe de variance qui module l'excitation. En outre, une enveloppe d'énergie de type triangulaire module le bruit comprenant la deuxième partie du spectre voisé dans le domaine temporel. Un filtre passe-haut de fréquence de coupure $f_m$ (fréquence de voisement maximum) est utilisé pour séparer la partie harmonique du bruit. Un schema simplifié des étapes d'analyse HNM est donné à la figure 2.5 [12]

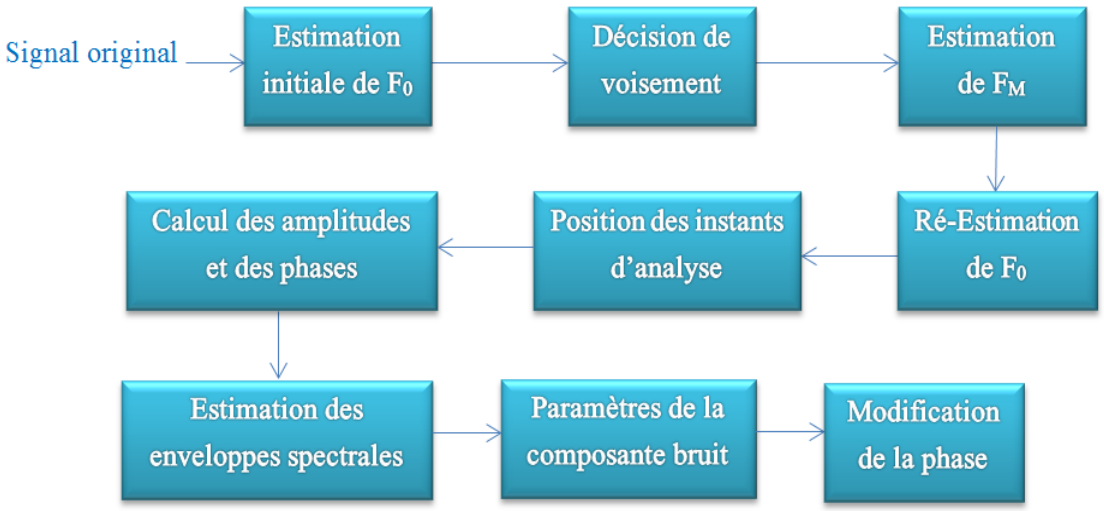


FIGURE 2.5 – Bloque diagramme simplifié des étapes d'analyse HNM

A chaque instant d'analyse, on dispose des éléments suivants : $f_0$ , $L$, amplitudes et phases et les coefficients (typiquement au nombre de 10) du filtre LPC pour la partie bruit. La partie harmonique et la partie bruit sont synthétisées séparément. La synthèse de la partie harmonique est réalisée de manière synchrone (pitch synchrones) et directement dans le domaine temporel comme une somme d'harmoniques . Les amplitudes et les phases de la composante harmonique, sont estimées via un critère des moindres carrés, et sont interpolées linéairement entre les trames successives. Seulement, les phases sont déroulées (unwrapped) avant d'appliquer l'interpolation. La synthèse de la partie bruit est effectuée comme suit : Un bruit Gaussien blanc de variance unitaire $b(t)$ est passé à travers le filtre $h(t)$ plusieurs fois par trame afin de s'assurer que les caractéristiques temporelles sont générées avec succès ; si la trame est voisée, le bruit est filtré en passe-haut avec une fréquence de coupure égale à la fréquence de voisement maximum $f_m$. Ensuite, pour s'assurer que le bruit est synchronisé avec la partie harmonique, il est modulé par une enveloppe temporelle.

Le signal synthétique voisé est obtenu en ajoutant les deux parties, c'est-à-dire partie harmonique synthétisée plus partie de bruit synthétisé. Le signal vocal synthétique non voisé est constitué simplement de la partie de bruit synthétisé. Le signal vocal synthétisé est calculé par recouvrement et addition de segments voisés et non voisés de la parole synthétique.

Il a été montré que la représentation R_HB a été appliquée avec succès dans plusieurs applications d'analyse-synthèse de la parole et présente beaucoup d'avantages par rapport aux représen-

tations de l'état de l'art [57, 58, 7].

Cependant, les représentations hybrides citées ci-dessus sont incapables de représenter des parties transitoires de la parole, telles que les consonnes plosives. Pour palier à ce problème, des modèles étendus ont été suggérés, généralement appelés representation sinusoïdale plus bruit plus transitoire.

### 2.2.4 Représentation Déterministique plus Transitoire plus Stochastique (R_DTS)

Afin de pouvoir modéliser les signaux très localisés en temps et caractérisés par des variations d'énergie très brusques, le modèle « sinusoïdal (ou harmonique) + transitoires + bruit » a été introduit dans [59, 60, 61]. Il s'agit ici des sons transitoires en parole (comme les consonnes plosives par exemple) ou des sons percussifs en musique qui sont caractérisés par des profils d'évolution temps-fréquence irréguliers sur des durées assez courtes (de l'ordre de la dizaine de millisecondes, voir moins).
Dans ce type de représentations hybrides, une phase de détection et de segmentation du signal de parole est utilisée pour traiter généralement de façon spécifique les zones transitoires du signal. Ainsi, des modèles spécifiques sont alors appliqués à ces zones transitoires, alors que le modèle sinusoïdes + bruit ou harmoniques + bruit est utilisé ailleurs pour les autres parties de la représentation hybride R_DTS.

Par exemple, Levine [59], propose de modéliser séparément les parties transitoires par une transformée de type "MDCT (Modified Discrete Cosine Transform)" ; et d'utiliser le modèle sinusoïdes + bruit pour les autres parties du signal sonore. On peut aussi citer le travail de [60] qui utilise un modèle à base de sinusoïdes amorties retardées.

Les méthodes d'analyse correspondantes aux représentations « sinusoïde plus transitoire plus bruit » sont très complexes et l'identification d'une trame transitoire n'est pas une tâche facile. Par conséquent, inclure le terme transitoire soit dans la partie déterministe, ou dans la partie stochastique est le choix le plus approprié.

## 2.3 Exemples de reconstruction sinusoïdale stationnaire

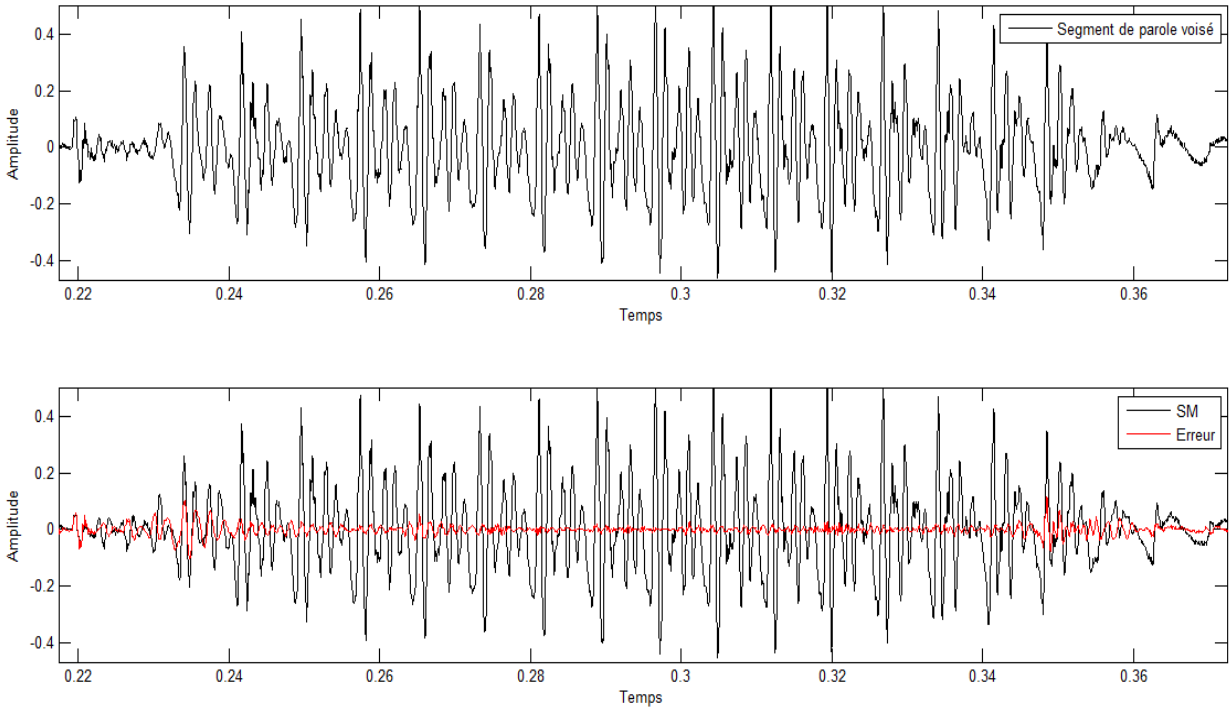


FIGURE 2.6 – Analyse-synthèse sinusoïdale avec erreur de reconstruction

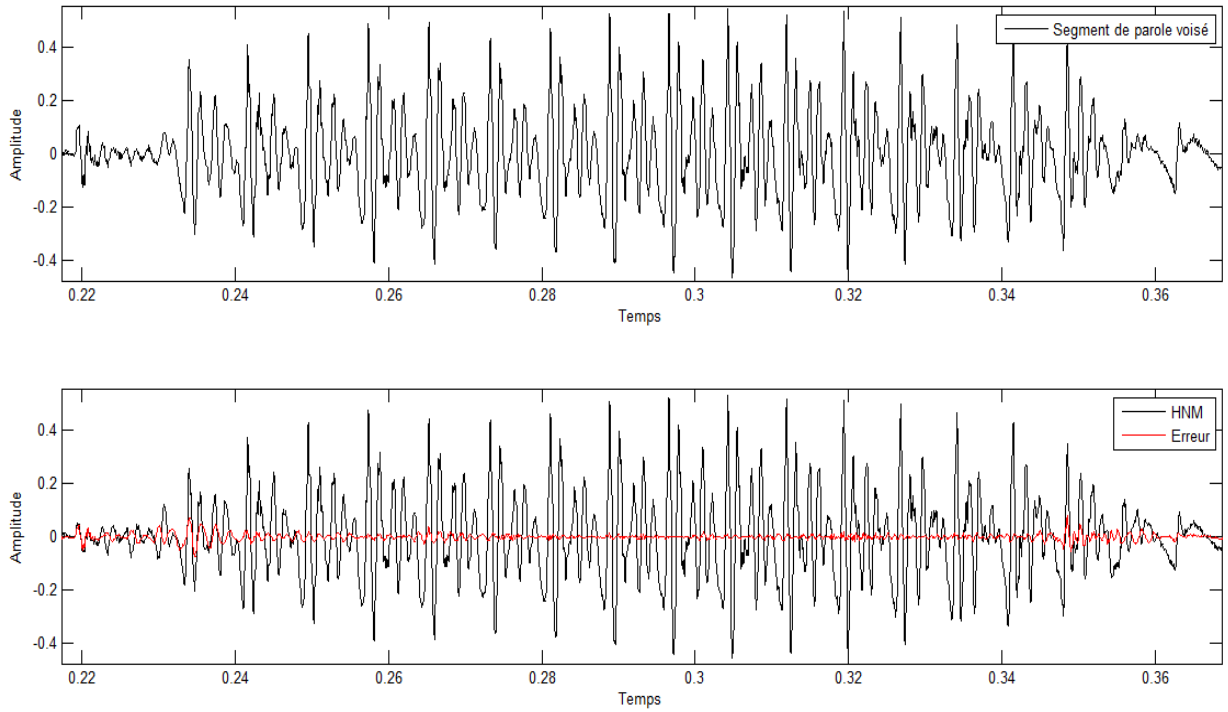


FIGURE 2.7 – Analyse-synthèse harmonic sinusoïdale avec erreur de reconstruction

Les figures 2.6, 2.7. montrent des exemples d'analyse-synthèse utilisant respectivement les deux représentations SM et HNM.

Un segment de parole voisé extrait d'une base de données arabe est analysé puis reconstruit en utilisant les modèles SM [10] et HNM [12] respectivement. Une fenêtre de Hamming est utilisée pour la pondération du segment analysé. Le nombre de composantes sinusoïdales est de 40. D'après les figures 2.6, 2.7, nous constatons que l'erreur de reconstruction produite par le modèle HNM est inférieure à celle du modèle SM. Ce qui confirme la haute qualité de synthèse vocale produite par le modèle HNM.

# Conclusion

Dans ce chapitre, nous avons décrit les représentations sinusoïdales stationnaires (uniformes ou hybrides) les plus utilisées dans les systèmes d'analyse-synthèse du son ou de la parole. Pour montrer la différence qui existe entre chaque type de représentation, leur processus d'analyse-synthèse a été également décrit. Des exemples expérimentaux ont été effectués sur des signaux de parole voisée (anglais et arabe) pour valider la performance de quelques types de ces représentations dans la reconstruction du signal et les résultats étaient très satisfaisants.

Cependant, pour améliorer la qualité de la reconstruction vocale et surtout pour les zones de paroles non stables, d'autres types de représentations ont été développées. Ces dernières vont être décrites en détail dans le prochain chapitre

# Chapitre 3

# Représentations sinusoïdales adaptatives du signal de la parole

Il a été montré que la reconstruction du signal de la parole utilisant les représentations sinusoïdales stationnaires décrites dans le chapitre précédent était de bonne qualité pour les types de signaux de parole relativement stables [10, 12]. Cependant, vu le caractère non stationnaire du signal de la parole, il a été suggéré de nouveaux modèles [14, 15, 16, 17, 18] qui s'adaptent aux caractéristiques locales du signal vocal pour donner une meilleure représentation et une haute qualité de reconstruction. Ces modèles sont connus sous le nom "adaptive Sinusoïdal Models (aSMs)".

Ce chapitre donc, fournit une brève description des représentations sinusoïdales adaptatives du signal vocal. Nous commençons par présenter le modèle quasi harmonique (QHM) [13] qui n'est pas adaptatif mais qui est à la base de tous les autres modèles adaptatifs. Ensuite, nous décrivons le modèle quasi harmonique adaptatif (aQHM) [14] suivi du modèle quasi harmonique adaptatif étendu (eaQHM) [15]. Enfin, des représentations sinusoïdales adaptatives hybrides [62, 63] et uniformes [16, 17, 18] seront exposées avec des exemples expérimentaux de reconstructions.

## 3.1 Analyse-synthèse basée sur les représentations sinusoïdales adaptatives

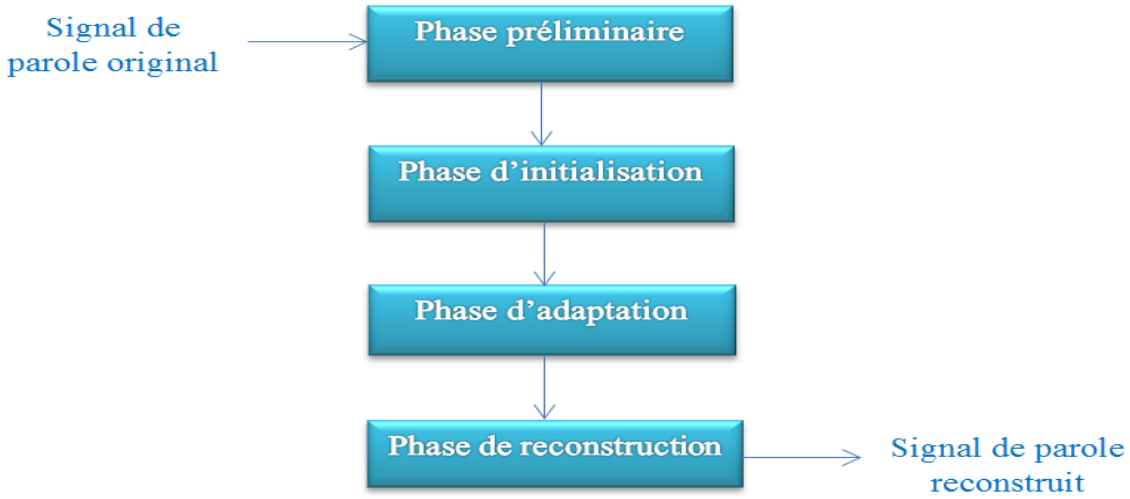


FIGURE 3.1 – Bloque diagramme simplifié d'un système d'analyse-synthèse aSMs

Un schéma général simplifié représentant les étapes d'analyse, d'adaptation et de synthèse utilisées par les modèles aSMs est représenté sur la figure 3.1. D'après la figure le système d'analyse-synthèse commence toujours par une analyse préliminaire qui consiste à l'estimation des fréquences du signal de la parole. Ces fréquences seront utilisées à l'étape d'initialisation par un modèle nommé QHM. Ce dernier, permet de corriger itérativement les erreurs d'estimation des fréquences obtenant ainsi une meilleure estimation des autres paramètres du modèle telles que les amplitudes et les phases. Cependant, le modèle QHM est toujours considéré comme modèle stationnaire. Ainsi, pour converger vers une représentation adaptative, autres modèles (aQHM ou eaQHM) seront utilisés à l'étape d'adaptation. A la sortie de cette étape, une meilleure estimation des paramètres instantanés du modèle adaptatif choisi sera obtenue. Enfin, à l'étape de synthèse, le signal de parole synthétique est reconstruit en utilisant les paramètres instantanés estimés avec une synthèse sinusoïdale additive.

## 3.2 Représentation Quasi harmonique (QHM)

Dans la représentation R_S, on peut représenter le signal vocal en utilisant les fonctions exponentielles au lieu des fonctions sinusoïdales de la manière suivante [62]

$$s(t) = A_0(t) + \sum_{l=1}^{L(t)} 2A_l(t)\cos(\phi_l(t)) = \sum_{l=-L(t)}^{L(t)} A_l(t)\exp(\phi_l(t)) \tag{3.1}$$

avec $A_l(t)$ est l'amplitude instantanée, $\phi_l(t)$ représente la phase instantanée du composant d'indice $l$. La dérivée de la phase instantanée donne la fréquence instantanée comme suit

$$f_l(t) = \frac{1}{2\pi}\frac{d\phi_l(t)}{dt} \tag{3.2}$$

L'hypothèse de la stationarité dans la représentation R_S, signifie que les amplitudes et les fréquences sont constantes, ce qui permet de donner la formule suivante

$$s_S(t) = (\sum_{l=-L}^{L} a_l \exp(j2\pi f_l t))w(t) \tag{3.3}$$

avec $L$ est le nombre locale des composantes , $f_l$, $a_l$ représentent respectivement la fréquence et l'amplitude locale.

Dans le cas où les fréquences des sinusoïdes sont des multiples entier de la fréquence fondamentale, on parle alors de la représentation sinusoïdale harmonique exprimée de la manière suivante

$$s_H(t) = (\sum_{l=-L}^{L} a_l \exp(j2\pi l f_0 t))w(t) \tag{3.4}$$

avec $f_0$ représente la fréquence fondamentale locale.

Dans la literature, il existe de nombreuses techniques pour l'estimation des paramètres inconnus de la représentation sinusoïdale ou sinusoïdale harmonique. Cependant ces techniques sont très sensible à l'estimation des fréquences. Une estimation erronée des fréquences conduit à des valeurs inexactes des autres paramètres de la représentation. Pour résoudre ce problème, l'utilisation de la représentation QHM [13] permet de corriger les erreurs d'estimation de la fréquence d'une manière itérative, ce qui donne une bonne estimation des amplitudes.

Le modèle QHM suggéré dans [13] qui est une version revisitée du modèle $HNM$ initialement proposé par Laroche et al. [38] est défini par

$$s_{QHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l)\exp(j2\pi \hat{f}_l t))w(t) \tag{3.5}$$

où $w(t)$, est la fenêtre d'analyse, $L$ est le nombre de composantes sinusoïdales (c'est-à-dire, l'ordre du modèle), $a_l$ est l'amplitude complexe, $b_l$ est la pente (slope) complexe et $\hat{f}_l$ est la fréquence d'analyse. Dans ce modèle, il est supposé qu'une estimation des fréquences réelles du signal vocal analysé est fournie à priori.

On assume que les vraie valeurs des fréquences du signal analysé sont inconnues et on définit l'erreur en fréquence par

$$\Delta f = f_l - \hat{f}_l \tag{3.6}$$

avec $f_l$ représentent les vraies fréquences.

Il à été montré dans [13] que le modèle QHM est capable de résoudre les erreurs en fréquences en projectant $b_l$ sur $a_l$ de la manière suivante

$$b_l = \rho_{1,l} a_l + \rho_{2,l} j a_l \tag{3.7}$$

où $ja_l$ représente le vecteur perpendiculaire à $a_l$. Les paramètres $\rho_{1,l}$ et $\rho_{2,l}$ sont obtenus par les équations suivantes

$$\rho_1 = \frac{\Re\{a_l\}\Re\{b_l\} + \Im\{a_l\}\Im\{b_l\}}{|a_l|^2} \tag{3.8}$$

$$\rho_2 = \frac{\Re\{a_l\}\Im\{b_l\} - \Im\{a_l\}\Re\{b_l\}}{|a_l|^2} \tag{3.9}$$

où les opérateurs $\Re\{.\}$ et $\Im\{.\}$ représentent, respectivement la parties réelle et imaginaire.

Il a été montré dans [13] que l'estimation de l'erreur de fréquence est donnée par

$$\Delta \hat{f}_l = \frac{\rho_2}{2\pi} = \frac{1}{2\pi} \frac{\Re\{a_l\}\Im\{b_l\} - \Im\{a_l\}\Re\{b_l\}}{|a_l|^2} \tag{3.10}$$

Les paramètres du modèle $a_l$ et $b_l$ sont estimés via une minimisation simple des $MC$ [62, 63] de la façon suivante :

Tout d'abord on définit un vecteur de paramètres comme suit :

$$x = \begin{bmatrix} a \\ b \end{bmatrix} \tag{3.11}$$

L'erreur est définie en temps discret par

$$\epsilon(a,b) = \sum_{l=-N}^{N} |s[n] - s_{qhm}[n]|^2 \tag{3.12}$$

$$= \sum_{l=-N}^{N} (s[n] - s_{qhm}[n])^*(s[n] - s_{qhm}[n]) \tag{3.13}$$

où $s[n]$ est le signal original, $s_{qhm}[n]$ est la représentation QHM définie par l'équation 3.5, $2N+1$ représente la taille de la fenêtre d'analyse .

En notation matricielle, si on sépare les valeurs de fenêtre des échantillons, l'équation 3.12 devient

$$\epsilon(a,b) = (Ws - Ws_{qhm})^H(Ws - Ws_{qhm}) \tag{3.14}$$

$$= W(s - s_{qhm})^H W(s - s_{qhm}) \tag{3.15}$$

$$= (s - s_{qhm})^H W^H W(s - s_{qhm}) \tag{3.16}$$

où $W$ est une matrice carrée ayant les valeurs de la fenêtre d'analyse dans sa diagonale, $s$ est un vecteur contenant les échantillons de signal d'origine, et H désigne l'opérateur Hermitien.

En passant à la notation matricielle dans le domaine discret, le modèle QHM devient

$$s_{QHM}[n] = \sum_{l=-L}^{L} (a_l + nb_l)\exp(j2\pi\hat{f}_l n/f_s) \tag{3.17}$$

$$s_{QHM}[n] = \sum_{l=-L}^{L} a_l \exp(j2\pi\hat{f}_l n/f_s) + \sum_{l=-L}^{L} nb_l \exp(j2\pi\hat{f}_l n/f_s) \tag{3.18}$$

$$s[n] = E_0 a + E_1 b = [E_0|E_1]\begin{bmatrix} a \\ b \end{bmatrix} = Ex \tag{3.19}$$

où

$$E_0 = (E_O)_{n,l} = \exp(j2\pi\hat{f}_l n/f_s) \tag{3.20}$$

$$E_1 = (E_1)_{n,l} = n(E_O)_{n,l} = n\exp(j2\pi\hat{f}_l n/f_s) \tag{3.21}$$

$$E = [E_0|E_1] \tag{3.22}$$

Par conséquent, la minimisation devient

$$\frac{\partial \epsilon(x)}{\partial x} = 0 \tag{3.23}$$

$$\frac{\partial}{\partial x}(s - Ex)^H W^H W(s - Ex) = 0 \tag{3.24}$$

La solution de cette équation est

$$x = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \tag{3.25}$$

Afin de corriger les fréquences un robust algorithme est suggéré dans [13]. Ce dernier est capable d'estimer de façon itérative les erreurs de fréquence.

Finalement, le signal peut être approximé par l'équation suivante

$$x_{Q\hat{H}M}(t) = (\sum_{l=-L}^{L} |a_l| \exp j(2\pi \hat{f}_l + \Delta \hat{f}_l)t + \hat{\phi}_l))w(t) \tag{3.26}$$

avec

$$\hat{\phi}_l = \angle \hat{a}_l \tag{3.27}$$

La représentation $QHM$ [13] assume toujours l'hypothèse de la stationarité des paramètres, ce qui n'est pas bien pour modéliser la non-stationarité qui existe dans le signal de la parole. Pour pallier à ce problème, l'utilisation des représentations adaptatives décrites dans les prochaines sections est nécessaire pour aboutir à une bonne analyse-synthèse du signal vocal vu sa nature non stationnaire.

## 3.3 Représentation adaptative quasi harmonique (aQHM)

Pantazis et al [14] ont proposé d'élargir le modèle QHM à une nouvelle représentation appelée modèle quasi harmonique adaptatif (aQHM) en projetant le signal sur des fonctions de phase variant dans le temps à l'intérieur de la fenêtre d'analyse comme suit

$$s_{aQHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l) \exp(j(\hat{\phi}(t + t_k) - \hat{\phi}(t_k))))w(t) \tag{3.28}$$

Avec $t_k$ représente le centre de la fenêtre d'analyse $w(t)$.

Comme dans la représentation QHM, les paramètres du modèle aQHM sont calculés en utilisant la méthode des MC :

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \tag{3.29}$$

où $W$ est une matrice carrée ayant les valeurs de la fenêtre d'analyse dans sa diagonale, $s$ est un vecteur contenant les échantillons de signal d'origine, et H désigne l'opérateur *Hermitien*.

la matrice $E$ est définie comme $E = [E_0|E_1]$ . Les sous-matrices $E_i, i = 0, 1$ ont des éléments donnés par

$$(E_O)_{n,l} = \exp(j(\hat{\phi}_l(t_n + t_i) - \hat{\phi}_l(t_i)) \tag{3.30}$$

et

$$(E_1)_{n,l} = t_n \exp(j(\hat{\phi}_l(t_n + t_i) - \hat{\phi}_l(t_i)) = t_n (E_O)_{n,l} \tag{3.31}$$

$\hat{\phi}(t)$ est la phase instantanée définie par

$$\hat{\phi}_l(t) = \hat{\phi}_l(t_i) + \int_{t_i}^{t+t_i} 2\pi f_l(\tau) d\tau \tag{3.32}$$

où $f_l$ représente la trajectoire fréquentielle estimée à l'aide d'une méthode d'estimation initiale des paramètres telle que l'approche du modèle QHM [13] et $t_i$ représente l'instant d'analyse .

Un algorithme itérative de décomposition (AM-FM)" Modulation d'amplitude et modulation de fréquence " qui permet d'estimer les paramètres instantanés de la représentation aQHM, à savoir l'amplitude $\hat{A}_l$ et la phase $\hat{\phi}_l(t)$ a été proposé dans [14].

La reconstruction du signal de la parole est considérée comme la somme des composantes (variables dans le temps) estimées par l'algorithme de décomposition (AM-FM)

$$\hat{s}(t) = \sum_{l=-L}^{L} \hat{A}_l(t) \exp(j\hat{\phi}_l(t)) \tag{3.33}$$

Ainsi une représentation quasi-harmonique adaptative de haute qualité de la parole voisée a été obtenue en utilisant un mécanisme de correction de fréquence et une adaptation de ses fonctions de base aux caractéristiques du signal d'entrée [14].

## 3.4 Représentation adaptative quasi harmonique étendue (eaQHM)

Dans la représentation aQHM, seule la phase est adaptée aux caractéristiques locales du signal de parole. Afin d'inclure l'adaptation d'amplitude locale, un nouveau modèle appelé modèle adaptatif quasi harmonique étendu (eaQHM) a été proposé dans [15]

$$s_{eaQHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l)\hat{\alpha}_l(t) \exp(j\hat{\phi}_l(t)))w(t) \tag{3.34}$$

$$\hat{\alpha}_l(t) = \frac{\hat{A}_l(t+t_k)}{\hat{A}_l(t_k)} \tag{3.35}$$

$$\hat{\phi}(t) = \hat{\phi}(t+t_k) - \hat{\phi}(t_k) \tag{3.36}$$

où $\hat{A}(t)$ et $\hat{\phi}(t)$ représentent l'amplitude instantanée et la phase instantanée, respectivement.

Comme dans la représentation $QHM$, les paramètres du modèle $eaQHM$ sont calculés en utilisant la méthode des $MC$ :

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (E_e^H W^H W E_e)^{-1} E_e^H W^H W s \tag{3.37}$$

où $W$ est une matrice carrée ayant les valeurs de la fenêtre d'analyse dans sa diagonale, $s$ est un vecteur contenant les échantillons de signal d'origine, et $H$ désigne l'opérateur *Hermitien*.

la matrice $E_e$ est définie comme $E = [E_0|E_1]$. Les sous-matrices $E_{ei}, i = 0, 1$ ont des éléments donnés par

$$(E_{eO})_{n,l} = \alpha_l(t_n) \exp j(\hat{\phi}_l(t_n + t_i) - \hat{\phi}_l(t_i) \tag{3.38}$$

et

$$(E_{e1})_{n,l} = t_n \exp(j(\hat{\phi}_l(t_n + t_i) - \hat{\phi}_l(t_i)) = t_n(E_{eO})_{n,l} \tag{3.39}$$

Le modèle eaQHM utilise ainsi une étape d'initialisation (QHM [13] par exemple) pour fournir une estimation initiale de l'amplitude $\hat{A}(t)$ et la fréquence $f_k(t)$ par interpolation linéaire et spline respectivement. Un schéma d'intégration de fréquence est utilisé pour estimer la phase $\hat{\phi}(t)$ [14]. Les paramètres $(a_l, b_l)$ sont estimés en utilisant un critère d'erreur des $MC$. Les paramètres du modèle (amplitudes et phases) sont mis à jour de manière itérative via un algorithme de décomposition (AM-FM)[14, 15].

La somme des composantes (variables dans le temps) estimées par l'algorithme de décomposition $AM - FM$ proposé dans [14] est utilisée ici pour reconstruire le signal vocal synthétique final utilisant l'équation 3.33.

Il a été montré dans [15] que la représentation eaQHM réalise une reconstruction très précise de la parole voisée, plus élevée que celle obtenue par la représentation aQHM [14] et par la représentation R_S [10]. Aussi, la représentation eaQHM a été appliquée avec succès pour modéliser une grande variété de sons de la parole tels que les sons percussifs, ou les sons d'arrêt [64] et les sons de parole non voisés [65].

## 3.5 Représentations sinusoïdales adaptatives hybrides

Les représentations aQHM et eaQHM sont principalement conçues pour modéliser des parties périodiques (voisées) du signal de la parole. Les parties non périodiques de ces modèles sont souvent représentées avec une composante aléatoire [66, 62] . Dans ce qui suit, nous allons décrire quelque représentations adaptatives hybrides.

### 3.5.1 Représentation adaptative quasi harmonique plus bruit (aQHNM)

La représentation aQHM a été appliquée avec succès à un système hybride d'analyse et de synthèse de la parole appelé modèle adaptatif quasi-harmonique plus bruit (aQHNM)[66, 62].

En prenant en compte les différentes sources qui constituent la parole, Pantazis et al. [66, 62] choisissaient de suivre une représentation hybride du signal vocal. Les représentations hybrides séparent le signal de la parole en une composante déterministe et une composante stochastique. La

composante déterministe modélise les caractéristiques quasi-périodiques de la parole tandis que la composante stochastique modélise les caractéristiques non périodiques de la parole.

La séparation du signal vocal, $s_{aqhnm}(t)$, en deux parties additives est donnée par

$$s_{aQHNM}(t) = D(t) + S(t) \tag{3.40}$$

où $D(t)$ désigne la partie déterministe alors que $S(t)$ désigne la partie stochastique. Les segments de parole voisés contiennent les deux parties tandis que la partie déterministe est nulle dans les segments non voisés.

Il a été proposé dans [66, 62] de modéliser la partie déterministe en utilisant le modèle $aQHM$ initialisé par le modèle $QHM$ comme dans l'algorithme de décomposition $AM - FM$ [14]. Tenant compte des caractéristiques variables du signal analysé, le modèle $aQHM$ est capable de répondre efficacement à la non-stationarité locale du signal de parole. Par rapport aux modèle $HNM$ [12]ou $SM$ [10], cette nouvelle approche réduit encore l'erreur dans l'estimation des paramètres sinusoïdaux, ce qui donne une représentation plus précise du signal.

La partie déterministe qui modélise les périodicités de segments vocaux exprimée en tant que somme de composantes sinusoïdales variant dans le temps comme à l'équation 3.1.

La fréquence instantanée est encore une fois donnée par l'équation 3.2.

La partie stochastique modélise les segments non voisés du signal vocal sous la forme d'un bruit Gaussien modulé en fréquence et en temps. Comme indiqué ci-dessus, la partie stochastique modélise toutes les informations des segments non voisés. Pour les segments voisés, la partie stochastique est définie comme le résidu entre le signal de parole et la partie déterministe reconstruite. Cependant, une partie déterministe ne peut pas entièrement représenter les périodicités, en particulier aux régions extrêmement non stationnaires du segment voisé, ainsi, le signal résiduel est filtré avec un filtre passe haut. En d'autres termes, cette étape de traitement affirme qu'en dessous d'une certaine fréquence, le signal voisé ne contient que des informations quasi périodiques. En résumé,

la partie stochastique est donnée par

$$S(t) = (s_{aqhnm}(t) - \hat{D}(t)) * F_p(t)) \tag{3.41}$$

où $\hat{D}(t)$ est la partie déterministe reconstruite alors que $F_p(t)$ est la réponse impulsionnelle d'un filtre passe-haut de phase zéro avec une fréquence de coupure $f_m$ .

La partie stochastique d'une trame $k$ est modélisée comme

$$\hat{S}^k(t) = e^k(t)[u^k(t) * q^k(t)] \tag{3.42}$$

où $u^k(t)$ désigne un processus de bruit Gaussien lié par un filtre $AR$ variable dans le temps avec une réponse impulsionnelle $q^k(t)$ alors que $e^k(t)$ est l'enveloppe temporelle. En ce qui concerne la modulation de fréquence, l'estimation du filtre $AR$ est effectuée par analyse $LP$ comme dans les segments non voisés. L'enveloppe temporelle (très importante pour la fusion des deux composants) est une enveloppe énergétique représentée comme somme des sinusoïdes.

L'idée derrière l'enveloppe d'énergie est de calculer la variation d'énergie de la composante stochastique et de la modéliser comme une somme de sinusoïdes de faible poids (ordre). L'enveloppe d'énergie de la partie stochastique est calculée par une moyenne locale de la partie stochastique absolue. Mathématiquement, l'enveloppe d'énergie est donnée par

$$e(t) = \int_{t-T_0}^{t+T_0} |S(\mu)| d\mu \tag{3.43}$$

où $To$ est 1 ms.

L'enveloppe temporelle pour la trame $k$ est ensuite approximé par une somme de sinusoïdes

$$\hat{e}^k(t) = \sum_{l=-L_e}^{L_e} ca_l^k \exp(j2\pi\xi_l^k t) \tag{3.44}$$

où $L_e$ est le nombre d'harmoniques qui est un petit entier, alors que les fréquences, $\xi_l^k$ et les amplitudes complexes, $ca_l^k$, sont calculées par sélection de pics du spectre de l'enveloppe temporelle comme dans le modèle sinusoïdal.

La composante stochastique est donc modélisée sous la forme d'un bruit Gaussien modulé dans le temps et modulé en fréquence. La modulation fréquentielle est réalisée par une modélisation

AR et une analyse LPC tandis que la modulation temporelle est réalisée par une enveloppe dans le domaine temporel. L'enveloppe dans le domaine temporel est très importante pour la fusion correcte des deux composants et (enveloppe basée sur l'énergie donne le meilleur résultat perceptuel). De plus, l'analyse de la partie stochastique peut être effectuée de manière asynchrone par rapport à la partie déterministe.

Dans l'étape de la synthèse, la partie déterministe est synthétisée sous la forme d'une somme variable dans le temps de sinusoïdes modulées en amplitude et modulées en fréquence. En effet, dans aQHM, l'interpolation trame par trame des paramètres est plus naturelle que la méthode OLA. Notez que cette méthode de synthèse est préférée par rapport à la méthode OLA car les trajectoires fréquentielles variant dans le temps étaient déjà utilisées par aQHM dans l'étape d'analyse [14].

D'autre part, la partie stochastique est synthétisée trame par trame en utilisant la méthode OLA. Pour chaque trame, un bruit blanc traverse le filtre AR pour obtenir la modulation de fréquence de la partie stochastique. Ensuite, l'enveloppe d'énergie est calculée et sa multiplication avec le bruit modulé en fréquence fournit la trame stochastique reconstruite.

Les tests d'écoutes ont montré [66, 62] que le signal reconstruit était indiscernable de l'original ce qui valide que la représentation $aQHNM$ de la parole était de haute qualité.

### 3.5.2 Représentation adaptative harmonique plus bruit (aHNM) et représentation adaptative quasi harmonique étendue plus bruit (eaQHNM)

Il a été présenté dans [63] deux systèmes hybrides d'analyse et de synthèse de la parole (eaQHNM, aHNM) basés respectivement, sur le modèle eaQHM [15] et sur le modèle adaptatif harmonique (aHM) [16, 17]. Le premier système (eaQHM) décompose le signal de parole voisé en des composantes modulées en amplitudes et en fréquences qui sont quasi-harmoniques tandis que le second système est inspiré de la théorie d'adaptation pour estimer avec précision la fréquence fondamentale qui est utilisée pour modéliser le signal vocal voisé.

Les parties non voisées de la parole sont représentées par un composant stochastique qui

est implémenté en tant que bruit Gaussien blanc modulé en temps et en fréquence pour les deux systèmes. Des exemples illustratifs sont donnés pour chaque modèle qui décrit leur performances dans le domaine temporel et fréquentiel [63].

Les deux modèles reposent sur un estimateur voisé / non voisé (V / NV) qui sépare les parties correspondantes de la parole. L'importance d'un tel estimateur est cruciale pour la performance des systèmes. Bien qu'un estimateur V / NV très simple soit utilisé dans leur travail, aucun artefact significatif n'est présent dans les formes d'onde de la parole resynthétisée. Selon la décision de l'estimateur, la trame de la parole sera modélisée soit par le modèle déterministe, soit par le modèle stochastique. Cela peut entraîner des problèmes dans les limites de la trame en raison de la fusion inappropriée entre différents modèles des trames adjacentes. Cependant, une décision binaire sur le voisement dans les trames est souvent erronée, puisqu'une trame peut être une trame transitoire, c'est à dire dans certains cas, elle ne peut pas être catégorisée sans équivoque comme étant voisée ou non.

Il a été suggéré dans [63] de laisser tomber l'estimateur V / NV, afin de réduire la complexité du système global et éliminer les éventuelles erreurs de classification de trames qui pourrait influencer les performances des systèmes hybrides. Une telle suggestion conduit à des systèmes uniformes à bande complète qui effectuent des décompositions AM-FM sur toute la longueur de la forme d'onde.

Il y a donc plusieurs raisons de suggérer une telle approche : tout d'abord, pour le signal de parole voisé, un certain nombre de systèmes hybrides s'appuient fortement sur une estimation précise de la fréquence $f_m$ qui divise le spectre de la parole voisée en une partie déterministique et une partie stochastique. L'estimation efficace de la fréquence $f_m$ est essentielle pour la performance du système Deuxièmement, comme il est décrit dans [16, 17],une telle fréquence n'est pas nécessaire du point de vue de la production de la parole.

## 3.6 Représentations sinusoïdales adaptatives uniformes

Ainsi, pour représenter les parties périodiques et non périodiques du signal de la parole de la même manière, des modèles adaptatives uniformes ont été proposés, à savoir, le modèle harmonique

adaptive (aHM) [16, 17] et le modèle eaQHM [18] qui seront décrit dans les sections qui suivent.

### 3.6.1 Représentation harmonique adaptative (aHM)

L'hypothèse derrière le modèle aHM est que le signal vocal peut être représenté comme suit

$$s_{a\hat{H}M}(t) = \sum_{l=-L}^{L} a_l(t) \exp(jl\phi_0(t)) \tag{3.45}$$

où $a_l(t)$ est une fonction complexe représentant à la fois l'amplitude et la phase instantanée et $\phi_0(t)$ désigne une fonction réelle décrite par

$$\phi_0(t) = \frac{2\pi}{f_s} \int_0^t f_0(\tau) d\tau \tag{3.46}$$

où $f_s$ est la fréquence d'échantillonnage, et $f_0$ représente la fréquence fondamentale qui est supposée être connue et peut avoir une erreur potentielle.

Dans la phase d'analyse, une séquence d'instants d'analyse est créée en utilisant la courbe $f_0(t)$ fournie auparavant. Autour de chaque instant d'analyse, une fenêtre de Blackman de 3 périodes de pitch locales est appliquée au signal vocal. Après cela, $\phi_0(t)$ est ensuite calculée au moyen d'une interpolation linéaire des fréquences $f_0^i$ et de l'intégration numérique de l'équation 3.46.

Afin d'obtenir les paramètres du modèle aHM, une méthode qui utilise le mécanisme de correction de fréquence du modèle aQHM est proposée dans [14]. Dans cette méthode, le modèle intermédiaire utilisé est défini par

$$s_{aQHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l) \exp(jl\phi_0(t)))w(t) \tag{3.47}$$

où $a_l$ et $b_l$ sont des valeurs complexes et $\phi_0(t)$ est encore définie par l'équation 3.46.

Afin d'avoir une estimation de $a_l$ et $b_l$, une minimisation par les MC est utilisée. Ces paramètres peuvent être utilisés pour estimer l'erreur de discordance(décalage) de fréquence (frequency mismatch error). Comme il est montré dans [16, 17], cette estimation, peut être utilisée à nouveau pour mettre à jour itérativement les valeurs de fréquence fondamentales $f_0$ et aussi le nombre de composants $L$. Un algorithme AIR (Adaptive Iterative Refinement) est alors suggéré [16, 17] pour

traiter la localisation des harmoniques hautes fréquences jusqu'à la fréquence de Nyquist.

Les paramètres instantanés du modèle aHM (amplitude $a_l$ et courbe de fréquence fondamentale $f_0$) sont obtenus par interpolation linéaire ou spline de leurs paramètres estimés aux instants de temps d'analyse calculés. Enfin, dans la phase de synthèse, le modèle aHM de l'équation 3.45 est utilisé pour générer chaque harmonique sinusoïdale à partir de ses paramètres estimés, harmonique après harmonique sans utiliser de fenêtre. Il a été montré dans [16, 17] que la représentation aHM peut pleinement répondre à la nature hautement non stationnaire des signaux de la parole et peut fournir une reconstruction de la parole de haute qualité.

### 3.6.2 Représentation adaptative quasi-harmonique étendue (eaQHM)

Inspiré par le modèle aHM uniforme de la section précédente, le modèle eaQHM initialement suggéré dans [15] a été encore amélioré à une nouvelle représentation appelée eaQHM à bande pleine [18]. Dans ce modèle, il a été supposé qu'un modèle harmonique initial converge successivement à la quasi-harmonicité.

Tout d'abord, une décomposition AM-FM à bande pleine est utilisée pour modéliser le signal de parole utilisant l'équation 3.1

La phase instantanée est donnée par l'équation 3.32.

On suppose qu'une estimation initiale de la fréquence fondamentale $f_0$ est fournie. Ensuite, une harmonicité à bande pleine est supposée afin d'obtenir une première estimation des amplitudes instantanées de tous les harmoniques. Par conséquent, initialement, un modèle harmonique simple est utilisé pour représenter une trame du signal vocal analysé. Afin d'estimer les paramètres du modèle, une minimisation par les MC est effectuée. Enfin, les paramètres $A_l(t)$ et $\phi_l(t)$ peuvent être initialement approximés en interpolant ces paramètres (amplitudes et fréquences) estimées sur des instants d'analyse successifs.

Pour converger vers une représentation adaptative quasi harmonique, le modèle eaQHM sug-

géré dans [15] est utilisé

$$s_{eaQHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l)\hat{A}_l(t)\exp(j\hat{\phi}_l(t)))w(t) \tag{3.48}$$

où $\hat{A}_l(t)$, $\hat{\phi}_l(t)$, et $\hat{f}_l(t)$ dénotent les paramètres estimés du modèle harmonique dans la phase d'analyse précédente.

Les paramètres $a_l$ et $b_l$ sont estimés par MC. Ces paramètres complexes sont utilisés pour former un terme de correction de fréquence pour chaque composante sinusoïdale. En utilisant ce terme de correction de fréquence, une estimation itérative des fréquences est effectuée. Cela conduit à une meilleure ré-estimation des composants instantanés du signal de la parole et le modèle initialement utilisé (c'est-à-dire le modèle harmonique) converge progressivement vers un modèle adaptatif quasi harmonique.

Dans la phase de synthèse, le signal vocal peut être reconstruit en utilisant l'équation 3.33.

Les paramètres de synthèse instantanés (amplitudes,fréquences et phases) sont calculés de la manière suivante :
$\hat{A}_l(t)$ est estimée par interpolation linéaire, $\hat{f}_l(t)$ est estimée par interpolation spline, et enfin, $\hat{\phi}_l(t)$ est estimée via une approche non-paramétrique basée sur l'intégration de la fréquence instantanée en utilisant l'équation 3.32.

Il a été montré dans [18] que le modèle eaQHM uniforme donne un signal synthétique de haute qualité, dépassant celles des autres représentations d'analyse-synthèse de la parole, telles que les représentations SM[10], HNM[12], aQHM[14] et aHM[16, 17].

## 3.7 Exemples de reconstructions sinusoïdales adaptatives

Le signal de la parole utilisé dans les exemples des figures 3.2, et 3.3 est un segment voisé arabe. L'analyse est faite en utilisant une fenêtre de Hamming. L'estimation initiale de la fréquence fondamentale est de 120 Hz. Cette valeur est utilisée par le modèle QHM. Les résultats d'estima-

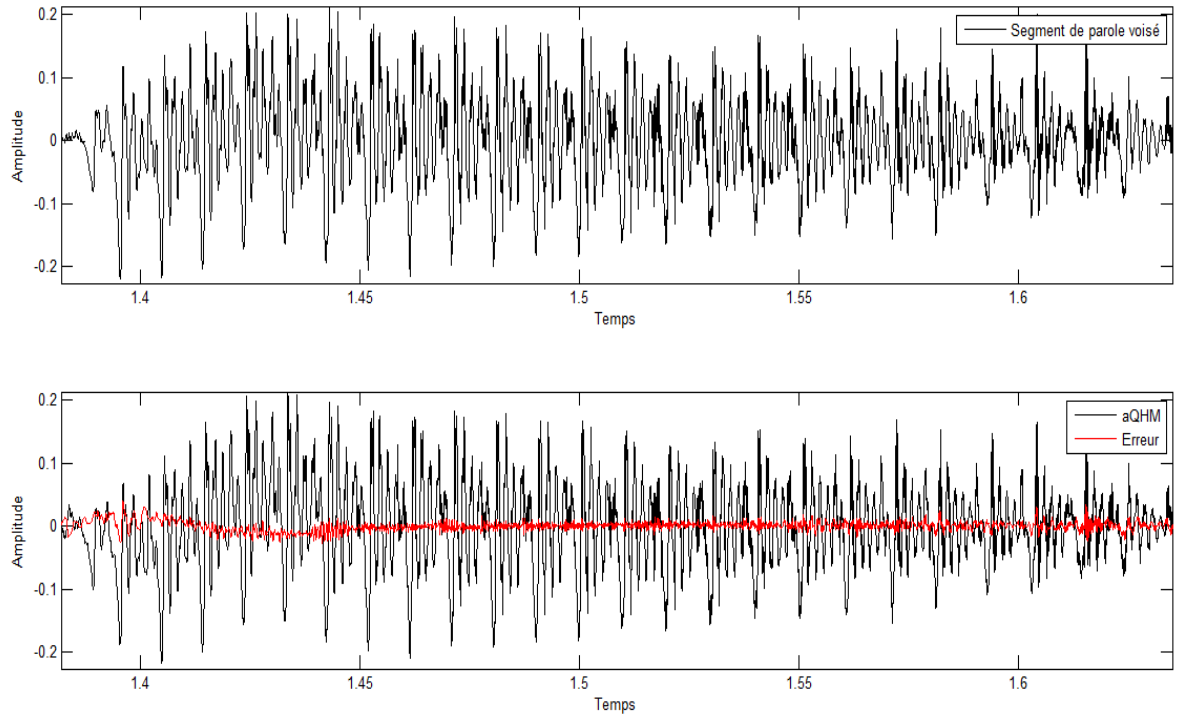


FIGURE 3.2 – Analyse-Synthèse et Erreur de reconstruction utilisant la représentation aQHM

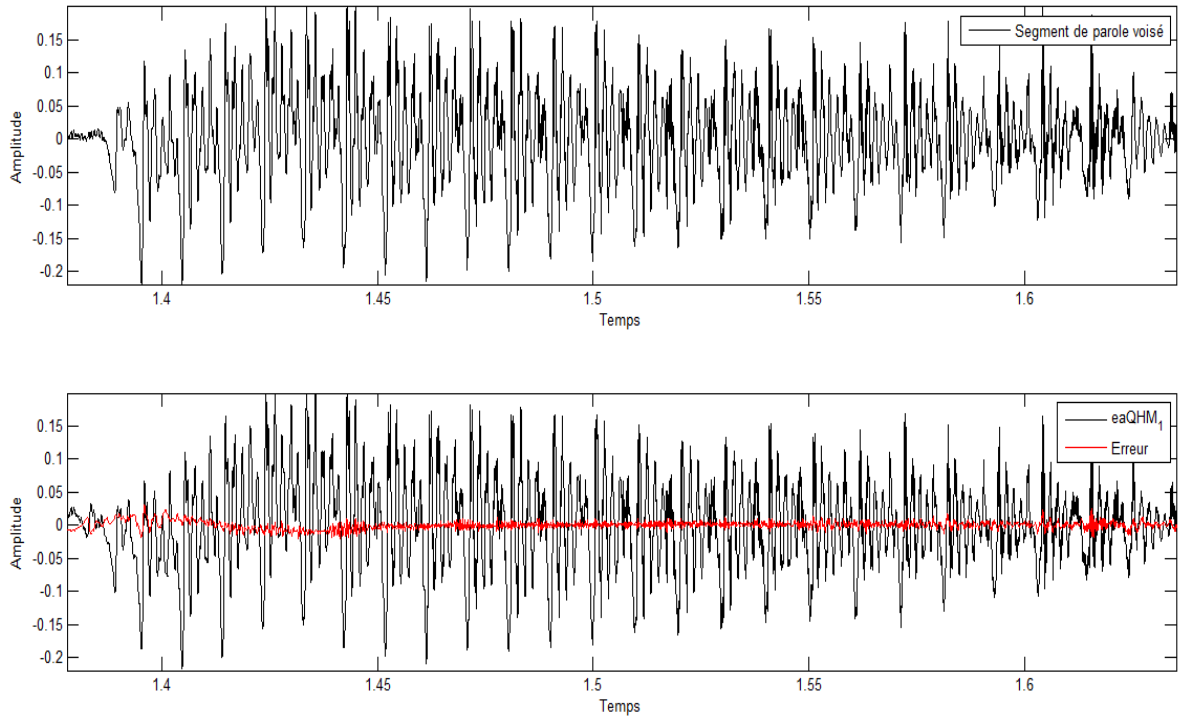


FIGURE 3.3 – Analyse-Synthèse et Erreur de reconstruction utilisant la représentation eaQHM

tion (fréquences, amplitudes et phases) obtenus par l'approche QHM seront utilisés comme valeurs initiales par les approches aQHM et eaQHM. Finalement, pour reconstruire le signal de parole, nous avons appliqué une synthèse sinusoïdale additive utilisant les paramètres instantanés estimés

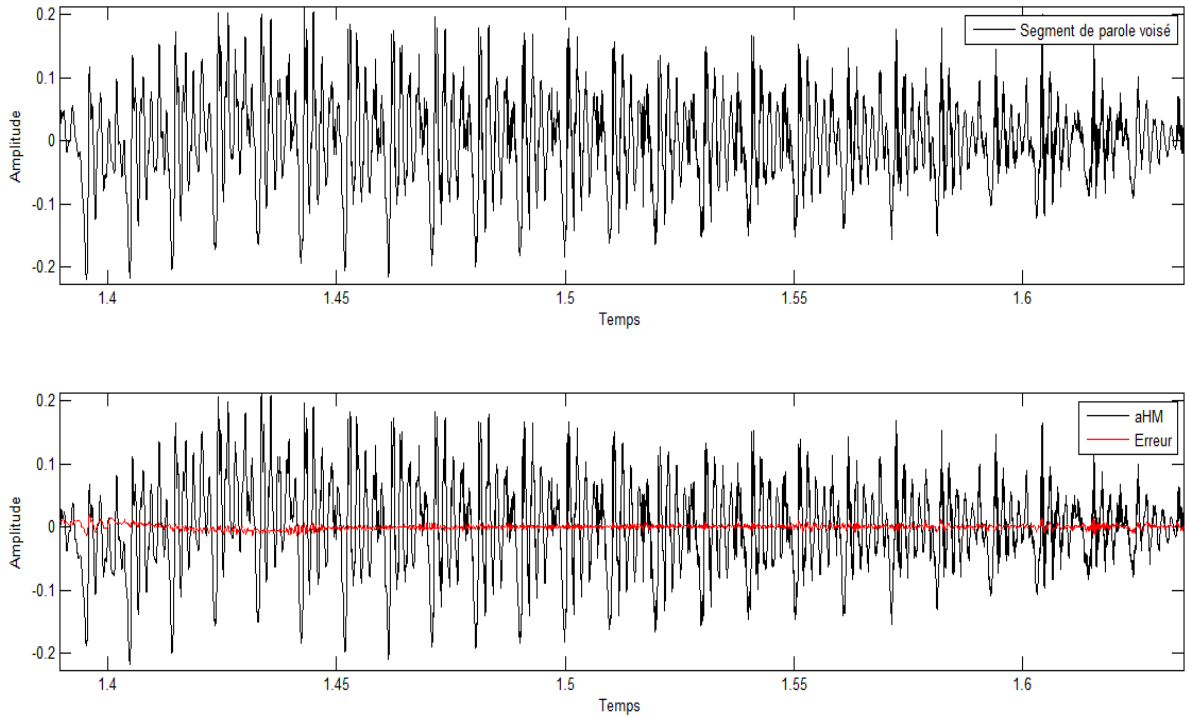


FIGURE 3.4 – Analyse-Synthèse et Erreur de reconstruction utilisant la représentation aHM

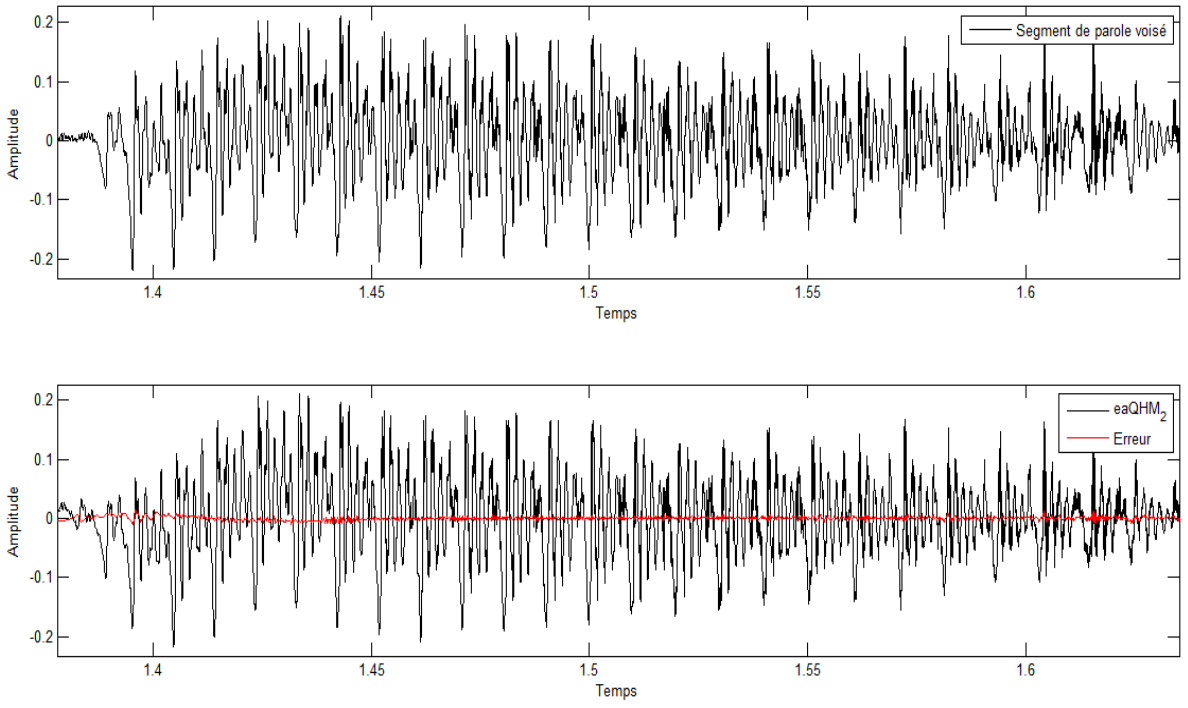


FIGURE 3.5 – Analyse-Synthèse et Erreur de reconstruction utilisant la représentation eaQHM uniforme

pour chaque modèle adaptative (aQHM ou eaQHM). D'après les résultats montrés aux figures 3.2, et 3.3, nous constatons bien que le modèle eaQHM produit une reconstruction meilleure que le mo-

dèle aQHM (l'erreur de reconstruction du modèle eaQHM est inférieure à celle du modèle aQHM).

Dans les figures 3.4 et 3.5 deux représentations adaptatives uniformes nommées aHM et eaQHM sont appliquées à un segment de parole voisé arabe. La fenêtre d'analyse utilisée est de type Blackman. Dans L'approche aHM, nous avons utilisé le modèle aQHM à l'étape d'adaptation, par contre, dans l'approche eaQHM, nous avons utilisé le modèle eaQHM. A l'étape de synthèse, nous avons utilisé les paramètres instantanés estimés par chaque approche avec une synthèse sinusoïdale additive. Les deux représentations donnent des résultats presque similaires vus les erreurs de reconstructions produites par chaque représentation.

# Conclusion

Dans ce chapitre, nous avons décrit les différents modèles sinusoïdaux adaptatifs récemment suggérés. Il a été montré qu'en appliquant un schéma adaptatif et un mécanisme de correction de fréquence, le signal vocal dans les modèles sinusoïdaux adaptatifs est représenté de manière très précise et compacte et la qualité du signal vocal synthétisé est largement améliorée avec une robustesse accrue par rapport aux représentations standards stationnaires de l'état de l'art. Par consequent, les représentations sinusoïdales adaptatives ont montré plus de potentiel pratique pour la reconstruction de la parole et ont offert une parole de haute qualité.

Basé sur la haute qualité de reconstruction obtenue par ces modèles sinusoïdaux adaptatifs, nous allons décrire dans le chapitre suivant une nouvelle représentation du signal de la parole nommée représentation sinusoïdale adaptative raffinée.

# Deuxième partie

# Contributions

# Chapitre 4

# Représentation sinusoïdale adaptative raffinée (R_aSR) du signal de la parole

Motivé par les performances obtenues pour la reconstruction du signal de la parole en utilisant les représentations aSMs (la représentation aHM [16, 17] et la représentation eaQHM [15, 18]), une nouvelle représentation sinusoïdale adaptative raffinée du signal vocal (R_aSR) est développée dans ce chapitre [19]. Nous allons montrer également dans ce chapitre, comment améliorer les étapes d'analyse et d'adaptation par apport aux travaux de l'état de l'art.

Pour valider l'efficacité et la robustesse de cette nouvelle représentation, plusieurs tests expérimentaux sont effectués sur un ensemble de signaux de parole voisée extraits à partir de deux différentes bases de données [67, 68].

## 4.1 Méthode proposée

Un schéma général décrivant notre méthode de représentation du signal de parole est illustré aux figures (4.1,4.2).

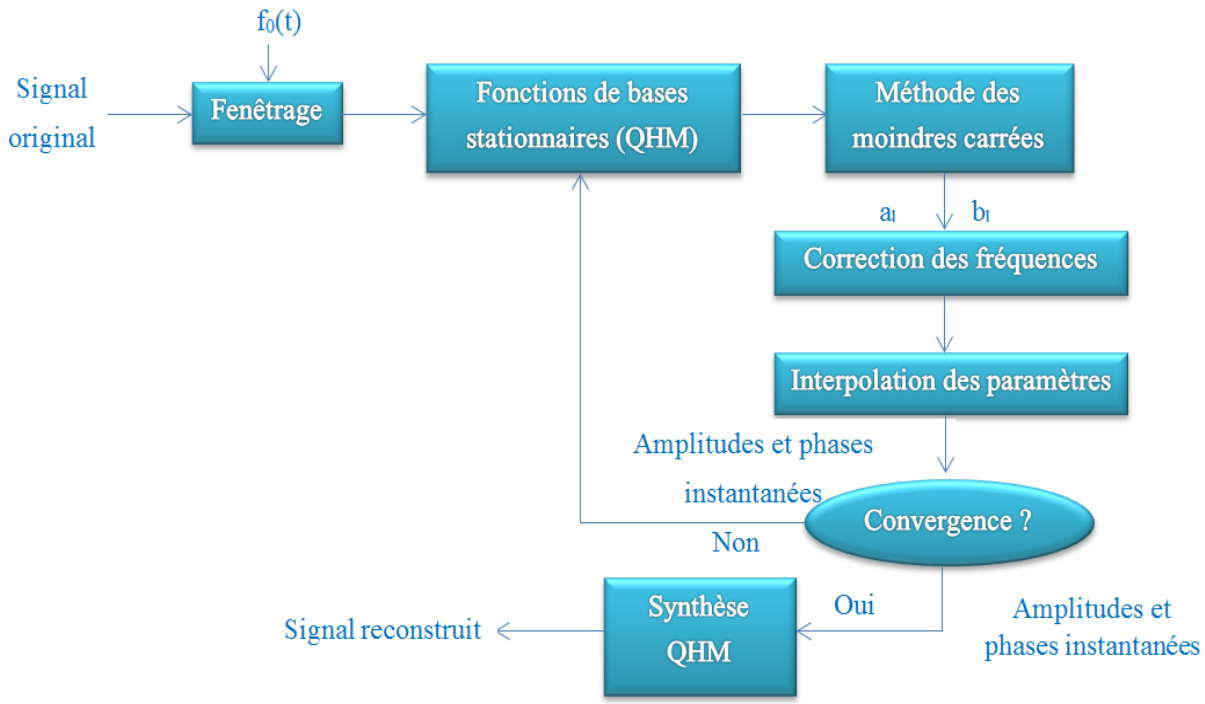


FIGURE 4.1 – Schéma simplifié de l'étape d'initialisation de la représentation R_aSR

Dans un premier temps, le modèle QHM [13] est utilisé dans la phase d'analyse afin d'obtenir une estimation initiale des paramètres instantanés de la nouvelle représentation R_aSR. Ensuite, dans l'étape d'adaptation, un schéma adaptatif combiné avec un mécanisme itératif de correction de la fréquence fondamentale est utilisé pour permettre une estimation robuste des paramètres du modèle (amplitudes, fréquences et phases).

En d'autre termes, le schéma adaptatif consiste à utiliser l'algorithme de décomposition AM-FM du modèle eaQHM [14, 15] pour une meilleur estimation des amplitudes et des fréquences instantanées. En ce qui concerne, le mécanisme itératif de correction de la fréquence fondamentale, nous avons utilisé, l'algorithme AIR du modèle aHM [16, 17].

Enfin, le signal vocal est reconstruit en tant que somme de ses composantes instantanées après

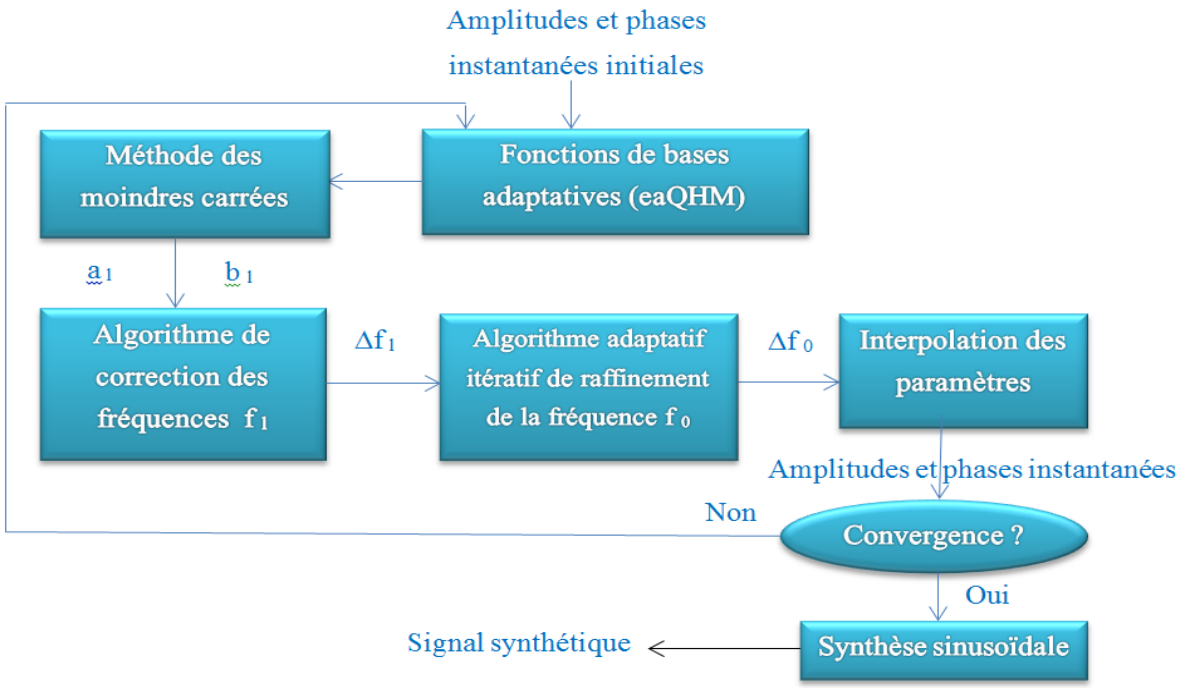


FIGURE 4.2 – Schéma simplifié des étapes d'adaptation de la représentation R_aSR

un mécanisme d'interpolation.

## 4.2 Analyse préliminaire

Vue que la parole est un processus non stationnaire, l'algorithme de décomposition AM-FM utilisé par le modèle eaQHM [15] ne peut pas être appliqué directement. Ainsi, il a été suggéré dans [62] d'ajouter un module qui effectue une estimation de fréquence fondamentale. Ayant une estimation de la fréquence fondamentale, l'étape d'initialisation et la définition des pistes de fréquence sont toutes les deux simplifiées. En effet, dans l'étape d'initialisation, QHM utilise comme fréquences initiales des multiples entiers de la fréquence fondamentale, tandis que les pistes de fréquences sont définies par le nombre d'harmoniques.

La fréquence d'échantillonnage varie entre 16 kHz et 48 kHz selon la base de données utilisée. Le nombre d'harmoniques $K = 40$.

Les étapes de l'analyse préliminaire sont les suivantes :

1. Segmenter le signal vocal en trames,

2. Séparer le signal vocal en ce qui est vraiment segment de parole et en ce qui ne l'en est pas,
3. Déterminer des régions voisées et non voisées de la parole,
4. Estimation de la fréquence fondamentale $f_0$ pour les régions voisées du signal de parole.

Nous avons besoin juste de donner la fréquence fondamentale estimée de la première trame voisée au début du segment de la parole voisée, $f_0(t_1)$, et ensuite faire l' hypothèses suivante : $f_l^0(t_1) = l.f_0(t_1)$

La fréquence fondamentale $f_0$ possède l'erreur potentielle suivante $\triangle f_0 = f_0 - \hat{f}_0$ où $\hat{f}_0$ représente une estimation de $f_0$.

Un ensemble de valeurs de temps d'analyse est ainsi calculé en utilisant l'ensemble de fréquences fondamentales estimées auparavant de la manière suivante :
$t_0 = 0$
$t_{i+1} = (f_0(t_i))^{-1} + t_i$

Autour de chaque instant d'analyse, le signal de parole est pondéré par une fenêtre de type Blackman et de longueur 03 pitch périodes (La limite supérieure de la taille de la fenêtre d'analyse est de 20 ms et la limite inférieure provient de la valeur minimum de la fréquence fondamentale qui est de 50 Hz).

## 4.3 Étape d'initialisation

Dans le but d'obtenir une estimation initiale des paramètres instantanés du modèle R_aSR, le fameux QHM [13] est utilisé en premier lieu comme suit :

$$s_{QHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l) \exp(j2\pi l f_0 t))w(t) \tag{4.1}$$

$f_0$ , étant donnée auparavant dans l'étape de l'analyse préliminaire, il reste à trouver les paramètres $(a_l, b_l)$ (amplitude et pente commplexe) du modèle en utilisant le critère des MC [62].

En utilisant lees paramètres estimés de $(a_l, b_l)$, une estimation du terme de la correction de la fréquence $\triangle f_l$ est donnée par

$$\triangle \hat{f}_l = \frac{1}{2\pi} \frac{\Re\{a_l\}\Im\{b_l\} - \Im\{a_l\}\Re\{b_l\}}{|a_l|^2} \tag{4.2}$$

où $\Re\{.\}$ et $\Im\{.\}$ désignent respectivement la partie réelle et la partie imaginaire.

En utilisant cette nouvelle estimation $\triangle \hat{f}_l$, les fréquences $f_l$ peuvent être mise à jour et le signal de parole peut être encore représenté par le modèle QHM, mais avec un nouvel ensemble de fréquences

$$\tilde{f}_l = \hat{f}_l + \triangle \hat{f}_l \tag{4.3}$$

Ainsi, Les étapes de l'algorithme itérative utilisé pour l'estimation des paramètres du modèle QHM sont les suivantes [62] :

---

**Estimation itérative des paramètres du modèle QHM**

---

1. **Initialisation**

   (a) Obtenir une estimation initiale des fréquences $\{\hat{f}_l\}$ pour $l = 1...L$,

   (b) Estimer $\{a_l, b_l\}$ à partir des valeurs initiales des fréquences estimées pour $l = 1...L$ en utilisant la méthode des MC

   ,

2. **Itérations**

   (a) Pour chaque composent :

      i. Estimer $\Delta \hat{f}_l$ en utilisant l'équation 4.2,

      ii. Mise à jour des fréquences : $\hat{f}_l = \hat{f}_l + \Delta \hat{f}_l$

      ,

   (b) Ré-estimer $\{a_l, b_l\}$ à partir des valeurs initiales des fréquences estimées pour $l = 1...L$ en utilisant la méthode des MC

   .

Finalement, le signal de la parole peut être reconstruit initialement en utilisant l'équation suivante

$$\hat{s}_{QHM}(t) = \sum_{l=-L}^{L} \hat{A}_l(t) \exp(j\hat{\phi}_l(t)) \tag{4.4}$$

où $\hat{A}_l(t)$ représente l'amplitude instantanée qui est estimée par interpolation linéaire de ces amplitudes estimées comme suit :

$$\hat{A}_l(t) = |a_l| \tag{4.5}$$

Par contre, la phase instantanée $\hat{\phi}_l(t)$ est égale à $l\hat{\phi}_0(t)$ où $\hat{\phi}_0(t)$ est obtenue par une interpolation de type spline de ces fréquences estimées et l'équation :

$$\hat{\phi}_0(t) = \frac{2\pi}{f_s} \int_0^t f_0(\tau) d\tau, \qquad \hat{\phi}_l(t) = l\hat{\phi}_0(t) \tag{4.6}$$

## 4.4 Adaptation

En utilisant le modèle QHM [13] à l'étape d'initialisation, le principe de la stationarité est toujours valide. Ainsi, pour converger vers une représentation adaptive en amplitude et en fréquence le modèle eaQHM [15] est utilisé comme suit :

$$s_{eaQHM}(t) = (\sum_{l=-L}^{L} (a_l + tb_l)\hat{A}_l(t) \exp(j\hat{\phi}_l(t)))w(t) \tag{4.7}$$

avec

$$\hat{A}_l(t) = \frac{\hat{A}_l(t + t_k)}{\hat{A}_l(t_k)} \tag{4.8}$$

et

$$\hat{\phi}(t) = \hat{\phi}(t + t_k) - \hat{\phi}(t_k) \tag{4.9}$$

où $\hat{A}(t)$ et $\hat{\phi}(t)$ représentent l'amplitude et la phase instantanées, estimées à la phase d'initialisation en utilisant l'approche du modèle QHM.

$t_k$ représente le centre de la fenêtre d'analyse

Il reste donc à estimer les paramètres $(a_l, b_l)$ à travers le critère des MC ce qui aboutit à une meilleur ré-estimation de l'amplitude instantanée $\hat{A}_l(t)$ et du terme de la correction de fréquence $\Delta\hat{f}_l$.

L'algorithme utilisé pour estimer ces paramètres est celui proposé dans [14, 15] et qui est donné par le pseudocode suivant :

---

**Algorithme itératif adaptatif de la décomposition AM-FM**

---

Une estimation initiale des amplitudes et phases est donnée par l'algorithme du modèle $QHM$

1. **Initialisation :**

   Donner une fréquence initiale $f_l^0(t_1)$

   Pour $k = 1, 2, \ldots, K$ (numéro de trame)

   (a) Calcul $(a_l^k, b_l^k)$ en utilisant $\hat{f}_l^0(t_k)$

   (b) Mise à jour de la fréquence $\hat{f}_l^0(t_k)$

   (c) Calcul $\hat{A}_l^0(t_k)$ et $\hat{\phi}_l^0(t_k)$

   (d) $f_l^0(t_{k+1})$ = $\hat{f}_l^0(t_k)$

   Fin de la boucle for

   Interpoler { $\hat{f}_l^0(t)$, $\hat{A}_l^0(t)$, $\hat{\phi}_l^0(t)$}

2. **Adaptation :**

   Pour $i = 1, 2, \ldots$

   (numéro d'adaptation) Pour $k = 1, 2, \ldots, K$

   (a) Calcul $(a_l^k, b_l^k)$ en utilisant $\hat{\phi}_l^{i-1}(t_k)$

   (b) Mise à jour de la fréquence $\hat{f}_l^i(t_k)$

   (c) Calcul $\hat{A}_l^i(t_k)$ et $\hat{\phi}_l^i(t_k)$

Fin de la boucle for $k$

Fin de la boucle for $i$

Interpoler $\{\hat{f}_l^i(t), \hat{A}_l^i(t), \hat{\phi}_l^i(t)\}$

---

En utilisant la nouvelle valeur du terme $\Delta\hat{f}_l$ obtenue à partir de l'algorithme ci-dessus, le terme de la correction de la fréquence fondamentale $\triangle\hat{f}_0$ peut être raffiné en utilisant l'algorithme AIR [16, 17].

Le terme de correction de fréquence $\Delta\hat{f}_l$ et relié au terme de correction de fréquence fondamentale $\triangle\hat{f}_0$ par la relation suivante :

$$\triangle\hat{f}_0 = \sum_{l=1}^{L} \frac{\Delta\hat{f}_l}{l} \tag{4.10}$$

Cette estimation peut en outre être utilisée pour mettre à jour le nombre d'harmoniques, L. Grace à la propriété de raffinement de fréquence utilisée dans l'approche utilisée par le modèle eaQHM, le terme $\triangle\hat{f}_0$ sera réduit progressivement et la valeur de $L$ augmente jusqu'à la fréquence limite de Nyquist. Le pseudocode de l'algorithme AIR [16, 17] est le suivant :

---

**Algorithme de raffinement iterative adaptatif - AIR**

---

— Créer une séquence de temps en utilisant la fréquence fondamentale $f_0(t)$

— Initialiser chaque $f_0^i$ = $f_0(t_i)$

— Initialiser chaque $L_i$ avec une petite valeur

Tant que $\exists i$ tel que $L_i.f_0^i < f_s/2$ faire

Pour chaque anchor $c$ faire

— Créer un segment de longueur égale à 3 fois le pitch autour $t_c$ en utilisant $f_0^c$

— Calculer $\phi_0(t)$ en utilisant l'équation 4.6 et l'interpolation de toutes les fréquences $f_0^i$

— Calculer $a_l^c, b_l^c$ en utilisant la solution MC

— Calculer $\triangle f_l$ et $\triangle f_0 = mean(\triangle f_l/l)$

— Corriger $fn_0^c = f_0^c + \triangle f_0$

Si $f_0^c K_c < f_s/2$ alors

Mettre à jour $K^c = 0.5Nw/|\triangle f_0|$

Fin si

Fin pour

Mettre $f_0^i = fn_0^i$

Fin tant que

---

Le terme anchor c, signifie les valeurs des paramètres instantanés estimés (amplitudes, fréquences et phases) aux instants d'analyses.

$N_w = minB_w, f_0$, avec $B_w$ représente la largeur de la bande du lobe principale de la fenêtre d'analyse.

A la convergence de l'algorithme AIR, des valeurs optimales de la fréquence fondamentale $\hat{f}_0$ seront obtenues et par conséquent, des valeurs de la phase instantanée $\hat{\phi}_l(t)$ seront raffinées.

## 4.5 Synthèse

Pour faire la reconstruction finale du signal de parole, une simple synthèse sinusoïdale est effectuée [16, 17] et qui consiste à la somme suivante :

$$\hat{s}_{R\_aSR}(t) = \sum_{l=-L}^{L} \hat{A}_l(t) \exp(j\hat{\phi}_l)(t) \tag{4.11}$$

Sachant que l'amplitude instantanée $\hat{A}_l(t)$ est calculée en utilisant l'interpolation linéaire. La fréquence instantanée $\hat{f}_0(t)$ est obtenue à travers une interpolation de type « spline ». Finalement, la phase instantanée $\hat{\phi}_l$ est obtenue en utilisant l'équation 4.6 et les valeurs estimées des fréquences.

## 4.6 Critère de convergence

Le mécanisme d'adaptation du modèle R_aSR continu jusqu'à la convergence. Le critère de convergence est relié à la quantité $SRER$ (Signal-to-Reconstruction Error Ratio) par la formule suivante [63] :

$$\frac{SRER_{n-1} - SRER_n}{SRER_{n-1}} < \epsilon \tag{4.12}$$

Tel que $\epsilon = 0.02$ , $n$ est l'indice de l'adaptation courante.

Le $SRER$ représente le rapport en décibel entre l'énergie du signal et l'énergie du résiduel et il est donné par la formule suivante

$$SRER = 20log_{10}\frac{std(s(t))}{std(s(t) - \hat{s}_{R\_aSR})} \tag{4.13}$$

avec $std$ représente la déviation standard, $s(t)$ est le signal de la parole original et $\hat{s}(t)$ est le signal synthétisé.

Le $SRER$ présente des valeurs positives élevé lorsque l'énergie du résiduel et plus faible de celle du signal original. Ce qui prouve une bonne précision dans l'estimation des paramètres du modèle.

Nous pouvons dire que l'algorithme d'adaptation converge lorsque la valeur du $SRER$ s'arrête d'augmenter.

## 4.7 Exemple de reconstruction utilisant la représentation R_aSR

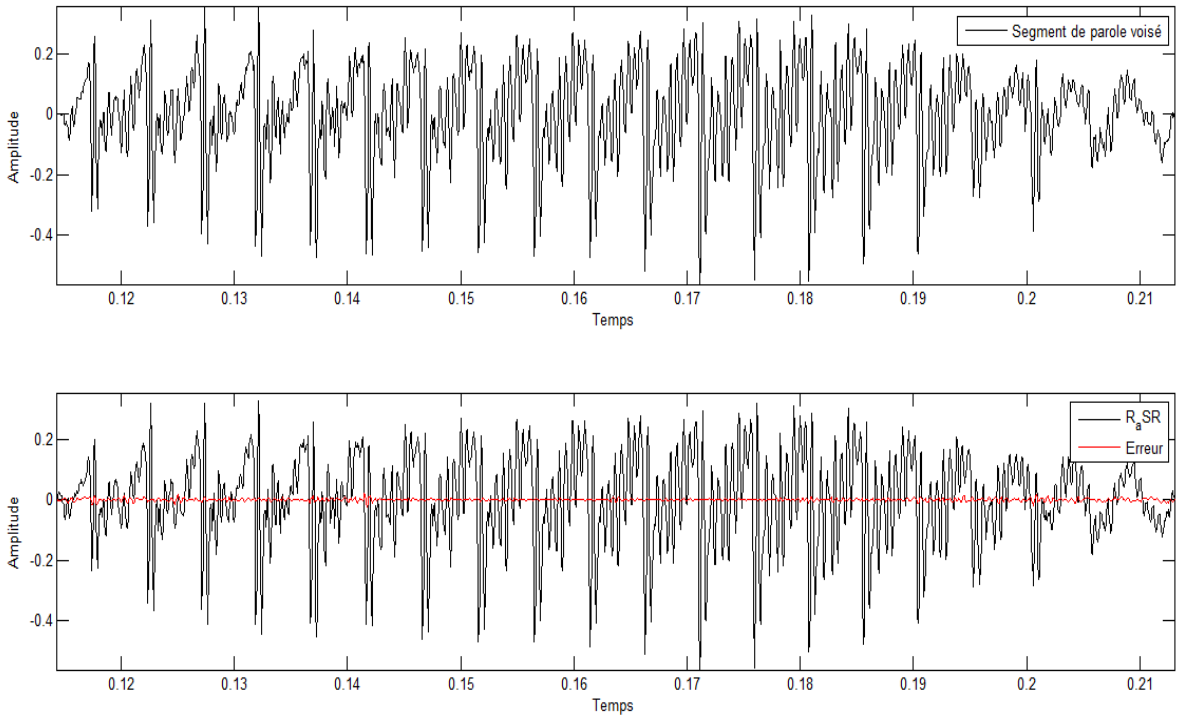


FIGURE 4.3 – Analyse-Synthèse et Erreur de reconstruction d'un segment voisé utilisant la représentation R_aSR

Un exemple d'analyse-synthèse d'un segment de parole voisé arabe utilisant notre nouvelle représentation R_aSR est illustré à la figure 4.3. Nous avons utilisé une fenêtre de type Blackman. A l'étape d'initialisation nous avons appliqué le modèle QHM pour obtenir une estimation initiale des paramètres instantanés. Ensuite, nous avons utilisé le modèle eaQHM pour raffiner l'estimation de ces paramètres instantanés. Enfin, une synthèse sinusoïdale additive utilisant les paramètres instantanés estimés est alors directement appliquée pour obtenir la signal de parole synthétisé comme montré à la figure 4.3.

D'après la figure, nous avons constaté que les deux signaux de parole, original et reconstruit, sont presque identiques et cela est confirmé par la faible erreur de reconstruction produite par notre représentation R_aSR.

# Conclusion

Dans ce chapitre nous avons décrit notre nouvelle représentation du signal de la parole nommée, représentation adaptative sinusoïdale raffinée (R_aSR). Vu que le signal de la parole est caractérisé par la propriété de la non stationnarité, notre nouvelle approche est basée sur un principe adaptatif en se basant sur les performances des modèles sinusoïdaux adaptatifs récemment suggérés. Par conséquent, nous avons amélioré les phases d'analyse, d'adaptation de notre nouvelle représentation par rapport à celles des représentations de l'état de l'art.

Les premiers résultats obtenus à partir de tests expérimentaux utilisant notre nouvelle représentation du signal de la parole sont très satisfaisants et pour confirmer cela, une application d'analyse-synthèse du signal vocal Anglais et Arabe basée sur notre nouvelle approche sera développée dans le chapitre suivant.

# Chapitre 5

# Application de la représentation sinusoïdale adaptative raffinée (R_aSR)

En se basant sur les performances des représentations aSMs et pour illustrer la performance de notre représentation R_aSR décrite dans le chapitre 4, nous allons développer dans ce chapitre un système d'analyse-synthèse de la parole basé sur la représentation R_aSR où nous allons analyser et reconstruire un ensemble de signaux de la parole voisée extrait de deux différentes bases de données [67, 68].

En titre de comparaison, nous allons comparer objectivement et subjectivement les résultats obtenus par notre système d'analyse-synthèse avec d'autres systèmes de l'état de l'art tel que : la représentation SM [10], la représentation HNM [12] et la représentation aHM [16, 17].

## 5.1 Applications des modèles sinusoïdaux adaptatifs

Comme mentionné dans le chapitre 3, les représentations aSMs [14, 15, 16, 17, 18] ont été utilisées avec succès dans plusieurs applications de traitement de la parole. Citons par exemple l'application de la représentation aQHM [14] à la reconstruction de la parole voisée où il a été prouvé que l'algorithme itératif de décomposition AM-FM basé sur la représentation aQHM peut être appliqué directement sur les signaux vocaux voisés et les résultats étaient très satisfaisants. Également, dans [66, 62], il a été développé un système d'analyse-synthèse vocal basé sur une représentation hybride de la parole, nommé aQHNM. Ainsi, le signal de la parole a été séparé en une partie déterministe et en une partie stochastique. La partie déterministe a été modélisée comme une somme de sinusoïdes variant dans le temps dont les composantes instantanées ont été estimées en utilisant le modèle aQHM et la partie stochastique a été modélisée en tant que bruit modulé en fonction du temps et de la fréquence. Les mêmes systèmes d'analyse-synthèse vocal basés sur une représentation hybride de la parole, ont été développés dans [63] mais cette fois-ci utilisant les représentations eaQHM et aHM pour la modélisation de la partie déterministique.

Bien que l'on ait démontré que les modèles hybrides fournissaient une flexibilité dans la manipulation et la resynthèse de la parole, il a été montré dans [16, 18] que l'adaptativité locale et l'harmonicité peuvent représenter perceptuellement toutes les parties du signal de la parole. Ainsi, des représentations uniformes à bandes pleines (full-band) telles que aHM [16, 17] et eaQHM [18] ont été utilisées avec succès dans des applications d'analyse-synthèse de la parole. Citons aussi l'application de la représentation eaQHM [64, 65] à la modélisation des signaux de la parole non voisés. Enfin, et en ce qui concerne les applications de la modification de la prosodie (time and pitch scale modification), nous pouvons citer les travaux [69, 70] basés sur la représentation aHM.

## 5.2 Système d'analyse-synthèse basé sur la représentation sinusoïdale adaptative raffinée du signal de la parole

Un schéma simplifié de notre système d'analyse-synthèse basé sur la représentation adaptative raffinée du signal de la parole est illustré à la figure 5.1.

Notre système comporte donc trois principales phases :

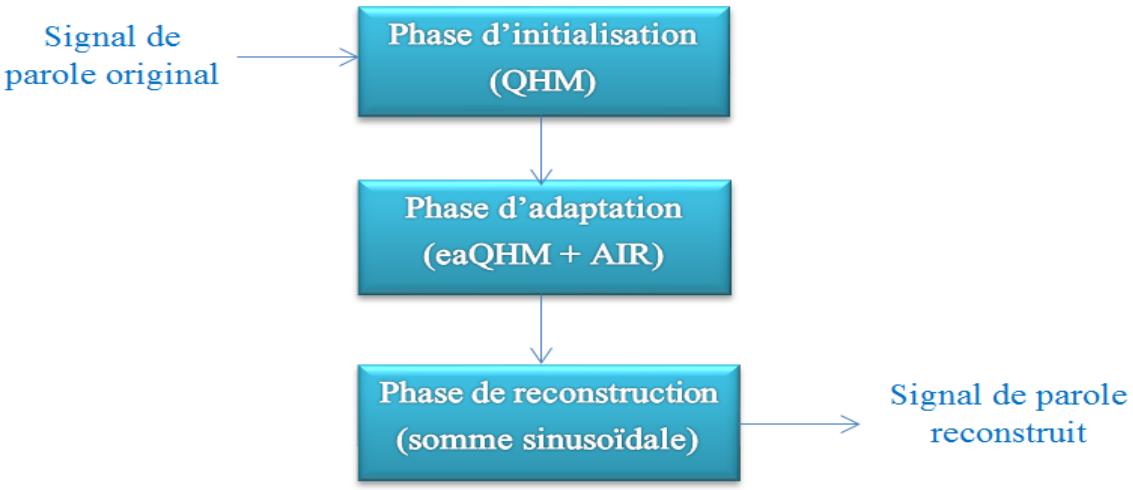


FIGURE 5.1 – Schéma simplifié du système d'analyse-synthèse basé sur la representation R_aSR

— Phase d'**initialisation** qui consiste à utiliser la représentation QHM [13] pour obtenir une première approximation des paramètres instantanés (amplitudes, fréquences et phases) de notre représentation

— Phase d'**adaptation** qui consiste à utiliser l'algorithme de décomposition AM-FM basé sur la représentation eaQHM [15] combiné avec l'algorithme AIR basé sur la représentation aHM [16] pour raffiner l'estimation des paramètres instantanés de notre représentation

— Phase de **reconstruction** qui consiste à une synthèse sinusoïdale additives du signal de la parole utilisant les paramètres instantanés estimés après un mécanisme d'interpolation.

## 5.3 Bases de données utilisées

Pour atteindre notre but, et afin de couvrir autant que possible la variabilité vocale, nous avons utilisé plusieurs signaux de la parole voisée provenant de deux différentes bases de données de paroles. Ces bases de données sont gratuites et accessibles au public sur Internet. La première base est constituée de signaux de parole en anglais ( ARCTIC speech database) [67] et la seconde est constituée de signaux de parole en arabe récemment développée par Halabi [68].

### 5.3.1 Base de données de parole Anglaise

Les bases de données **CMU_ARCTIC** [67] ont été construites à l'Institut des technologies des langages de l'Université Carnegie Mellon. Ces bases de données ont été conçues pour la recherche en synthèse vocale. Les bases de données se composent d'environ 1150 énoncés choisis avec soin. Les énoncés de la parole anglaise comprennent des locuteurs mâles / femelles , anglais/américains ainsi que d'autres locuteurs. La fréquence d'échantillonnage des énoncés sélectionnés est de 16 kHz et la durée de chaque segment vocal est d'environ 0,30 seconde.

### 5.3.2 Base de données de parole Arabe

La base **Arabic speech corpus** est un corpus de parole de l'arabe moderne standard (MSA) principalement construit pour être utilisé dans les applications de synthèse vocale. Le corpus contient des transcriptions phonétiques et orthographiques de plus de 3,7 heures de parole MSA ( 1813 fichiers .wav). Le Corpus a été construit dans le cadre d'un projet de doctorat par **N. Halabi** [68] à l'Université de Southampton. La fréquence d'échantillonnage des énoncés sélectionnés est de 48 kHz et la durée de chaque segment vocal est d'environ 0,30 seconde.

## 5.4 Classification voisée / non voisée / silence

Un algorithme de Classification voisée / non voisée / silence (V / NV / S) est exécuté en tant qu'étape de pré-traitement [62, 63]. La détection V / UV est effectuée dans une procédure trame par trame, avec une taille de trame de 30 ms avec un pas de chevauchement de 5 ms.

L'énergie de chaque trame est calculée et si elle est supérieure à un certain seuil, alors il est assigné comme parole. Sinon, cette trame est considérée comme un silence. Une fois la trame a été marqué comme parole, la classification voisée / non voisée, doit verifier certaines conditions pour confirmer que la trame considérée est voisée ou pas.

## 5.5 Estimation de la fréquence fondamentale

La fréquence fondamentale $f_0$ d'un signal (son ou pas) n'existe que pour les signaux périodiques (parole voisée par exemple), et elle est définie comme l'inverse de la période $T_0$ du signal.

Plusieurs algorithmes d'estimation de la fréquence fondamentale sont apparues dans la littérature. La précision et la robustesse de l'estimation de la fréquence fondamentale sont donc très importantes puisqu'elle représente le point de départ de bon nombre de méthodes d'analyse et synthèse vocale. Traditionnellement, il y a deux types d'algorithmes d'estimation de la fréquence fondamentale : les algorithmes basés sur la représentation fréquentielle (le spectre) du signal, et les algorithmes basés sur la représentation temporelle ( fonction d'auto-corrélation) du signal [4]. Puisque nos représentations utilisent les mécanismes d'adaptation, toute approche robuste peut être utilisée pour l'estimation initiale de la fréquence fondamentale.

Pour notre cas, nous avons utilisé l'estimateur de fréquence fondamentale récemment proposé nommé SWIPE (Sawtooth Waveform Inspired Pitch Estimator) [71]. Ce dernier est un estimateur de fréquence fondamentale développé pour le traitement de la parole et la musique. SWIPE estime le pitch comme étant la fréquence fondamentale d'une forme d'onde en dents de scie dont le spectre correspond le mieux au spectre du signal d'entrée.

## 5.6 Exemple de reconstruction de la parole arabe

Un exemple de reconstruction vocale voisé est présenté dans la figure 5.2 comme suit :
Un segment vocal voisé original arabe extrait de notre base de données [68] est montré dans le panneau $a$ et quatre signaux de parole reconstruits avec leurs erreurs de reconstruction correspondantes sont affichées dans les panneaux $b - e$.

## 5.7 Tests d'évaluations

Pour illustrer la performance de la représentation adaptative suggérée, Deux tests d'évaluations différents ont également été réalisés. Dans le premier test d'évaluation, le signal de parole original et reconstruit sont comparés objectivement en utilisant le rapport signal sur erreur de reconstruction (SRER) qui représente une mesure de la précision de la modélisation. Cependant, dans le deuxième test d'évaluation, un test d'écoute est effectué afin de comparer, le signal de parole original et reconstruit subjectivement en utilisant les mesures MOS (Mean Opinion Score) [72].

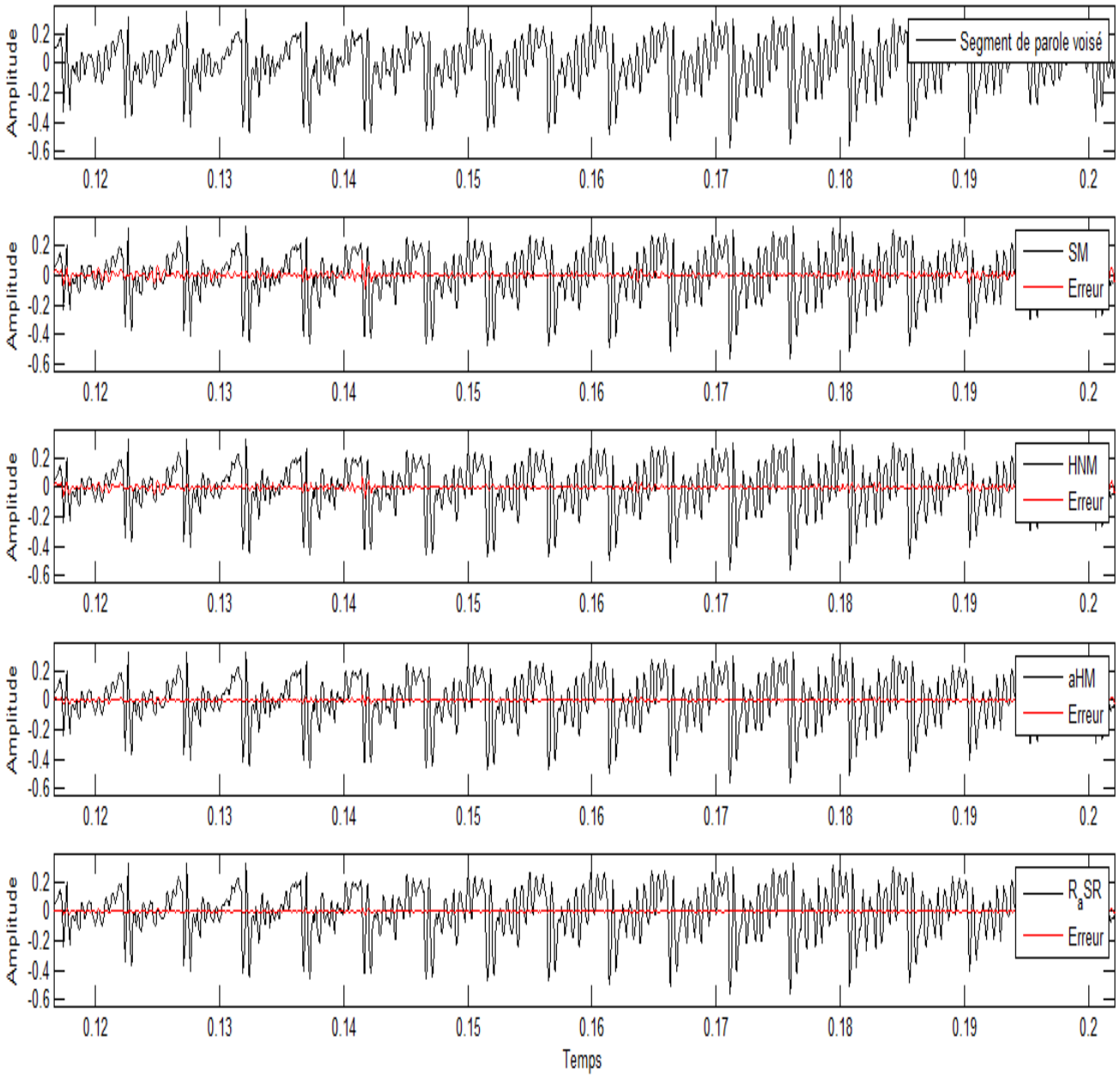


FIGURE 5.2 – Exemples de reconstruction vocale voisé arabe. (a) Segment de parole voisé (b) Reconstruction SM et erreur (c) Reconstruction HNM et erreur (d) Reconstruction aHM et erreur (e) Reconstruction R_aSR et erreur

### 5.7.1 Test d'évaluation objective

Dans le test d'évaluation objective, une comparaison mathématique entre le signal vocal original et le signal reconstruit est donc faite en mesurant le SRER [62], qui représente une distance numérique (c'est-à-dire une mesure de distorsion) entre les deux signaux comparés. Un ensemble

TABLE 5.1 – SRER moyens de l'évaluation objective

| Modèle | SRER |
|---|---|
| SM | 21,33 |
| HNM | 27,75 |
| aHM | 39,85 |
| R_aSM | 40,01 |

de signaux de parole voisé extrait de nos bases de données (anglaise et arabe) sont analysés et reconstruits en utilisant les modèles : SM [10], HNM [12], aHM [16, 17] et (R_aSR), respectivement, afin d'illustrer la capacité de chaque modèle à capturer des informations à partir d'un signal vocal original.

La mesure SRER correspondante à chaque représentation est calculée en utilisant l'équation suivante :

$$SRER = 20 \log_{10} \frac{std(s(t))}{std(r(t))} \tag{5.1}$$

où $std$ est la déviation standard, $s(t)$ est le signal original et $r(t)$ représente le résiduel entre le signal original et le signal reconstruit.

Rappelons que le $SRER$ représente le rapport en décibel entre l'énergie du signal et l'énergie du résiduel et qu'il présente des valeurs positives élevé lorsque l'énergie du résiduel et plus faible de celle du signal original. Ce qui prouve une bonne précision dans l'estimation des paramètres du modèle.

La table 5.1, résume les résultats du SRER moyen (en dB) pour chaque représentation.

Les tables 5.2, 5.3, 5.4, 5.5 résument les résultats du SRER moyen (en dB) pour chaque représentation et pour chaque type de signal de parole voisé (voyelle ou consonne / arabe ou anglaise).

TABLE 5.2 – SRER moyens de l'évaluation objective des voyelles arabes courtes

| Modèle | /a | /i | /u |
|---|---|---|---|
| SM | 21,33 | 23,11 | 22,09 |
| HNM | 27,75 | 29,13 | 28,22 |
| aHM | 39,85 | 40,12 | 41,05 |
| R_aSM | 43,21 | 42,33 | 42,76 |

TABLE 5.3 – SRER moyens de l'évaluation objective des voyelles anglaises

| Modèle | /a | /i | /u |
|---|---|---|---|
| SM | 24,13 | 22,53 | 23,17 |
| HNM | 29,15 | 30,09 | 29,87 |
| aHM | 40,35 | 39,90 | 41,66 |
| R_aSM | 41,09 | 40,67 | 39,96 |

TABLE 5.4 – SRER moyens de l'évaluation objective des consonnes arabes

| Modèle | /D | /z | /G |
|---|---|---|---|
| SM | 19,13 | 18,65 | 19,23 |
| HNM | 25,95 | 26,76 | 25,83 |
| aHM | 38,09 | 37,59 | 37,90 |
| R_aSM | 39,01 | 38,89 | 39,33 |

TABLE 5.5 – SRER moyens de l'évaluation objective des consonnes anglaises

| Modèle | /b | /v | /d |
|---|---|---|---|
| SM | 18,85 | 19,15 | 19,77 |
| HNM | 26,14 | 25,98 | 26,33 |
| aHM | 37,88 | 38,64 | 37,79 |
| R_aSM | 38,95 | 39,71 | 39,01 |

### 5.7.2 Test d'évaluation subjective

Dans le test d'évaluation subjective et selon la recommandation UIT-R BS [72], les signaux de parole originaux et reconstruits sont comparés par un groupe d'auditeurs à qui nous avons demandé de noter la qualité vocale synthétique en utilisant l'échelle suivante (5 : excellent, 4 : Bon, 3 : Passable, 2 : médiocre, 1 : Mauvais). Le score moyen résultant obtenu auprès de tous les auditeurs est appelé mesures du score moyen d'opinion (MOS). L'enregistrement vocal voisé original suivi des signaux vocaux reconstruits de chaque modèle, sont présentés comme modèle 1, modèle 2, modèle 3 et modèle 4 dans un ordre aléatoire et les participants aux tests d'écoute sont demandés d'écouter et d'évaluer la qualité perçue de chaque discours resynthétisé. La figure 5.3 présente les résultats de cette évaluation subjective en termes de mesures MOS, montrant ainsi la qualité de la reconstruction perçue pour chaque modèle utilisé par rapport au signal vocal original.

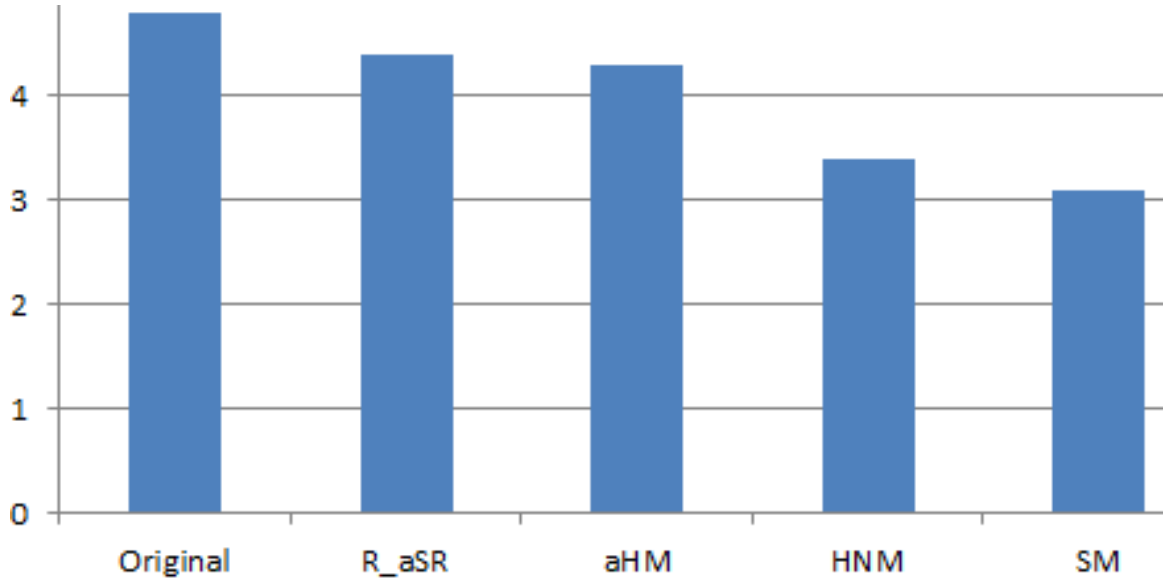


FIGURE 5.3 – Résultats du test d'écoute en termes de mesures MOS

## 5.8 Discussion

Notre représentation a été évaluée en faisant des comparaisons avec le modèle adaptatif aHM et les deux modèles stationnaires, à savoir SM et HNM, en utilisant un sous-ensemble d'enregistrements vocaux de bases de données vocales en arabe et en anglais. Certaines informations ont d'abord été données sur le schéma d'analyse-synthèse de chaque modèle. Ensuite, chaque énoncé vocal a été analysé et synthétisé en utilisant tous les modèles comparés. Enfin, des évaluations nu-

mériques (c'est-à-dire des métriques) telles que les mesures SRER et MOS ont été calculées afin d'évaluer la qualité et la transparence du signal de parole reconstruit de chaque modèle.

Comme nous pouvons l'observer à partir de l'exemple de reconstruction de la figure 5.2, les erreurs de reconstruction sont réduites en utilisant les modèles R_aSR et aHM, ce qui confirme la robustesse de la représentation proposée dans l'estimation des paramètres instantanés du modèle.

D'après les résultats décrits dans les tableaux 5.1, 5.2, 5.3, 5.4, 5.5, nous pouvons voir que R_aSR et aHM ont fourni des mesures (SRER) élevées par rapport à SM et HNM. Par conséquent, la reconstruction de haute qualité de la parole a été prouvée en utilisant notre représentation (R_aSR).

Évaluer quel signal reconstruit était perceptivement plus proche du signal vocal d'origine est le but des tests d'évaluation de l'écoute, et en général, les participants ont reconnu que la reconstruction fournie par le modèle R_aSR est naturelle comme le montre la figure 5.3. De même, le modèle aHM a fourni une qualité perçue transparente par rapport aux modèles SM et HNM.

# Conclusion

Un système d'analyse-synthèse basé sur notre nouvelle représentation sinusoïdale adaptative raffinée décrite dans le chapitre 4 a été développé dans ce chapitre. Pour atteindre notre objectif, nous avons utilisé deux bases de données différentes constituées d'énoncés en langue arabe et en langue anglaise. Un module d'estimation de la fréquence fondamentale a été présenté en premier, ensuite les étapes de réalisations de notre système ont été décrites en détails. Pour l'évaluation globale de notre nouveau système appliqué à la synthèse de signaux de parole voisés, nous avons utilisé des tests d'évaluations subjectives et objectives. Une discussion autour des résultats obtenus a été donnée et dans l'ensemble le système d'analyse-synthèse produit une qualité de synthèse vocale arabe très satisfaisante.

# Conclusions et futures perspectives

# Conclusions

L'objectif principal de cette thèse est de développer une représentation du signal de la parole relativement simple, flexible, de haute qualité et robuste pour les systèmes d'analyse-synthèse vocale. Dans cette direction, diverses représentations de signaux vocaux utilisées dans les applications d'analyse/synthèse vocale ont été discutées dans cette thèse. L'accent est mis sur les représentations sinusoïdales adaptatives.

Les représentations sinusoïdales stationnaires, telles que le modèle sinusoïdale (SM) ou le modèle harmonique plus bruit (HNM), fonctionnent bien sous l'hypothèse de la stationnarité à court terme de la parole. C'est-à-dire que les sinusoïdes qui représentent la parole ont des amplitudes et des fréquences constantes pour une fenêtre d'analyse de courte durée. Cependant, Il a été déjà montré que ce n'est pas le cas dans les signaux de paroles, où il y a des changements rapides et non linéaires d'amplitude et de fréquence pendant de courts intervalles de temps et il est essentiel de capturer ces fluctuations de courte durée pour avoir une analyse-synthèse de la parole de haute qualité. Dans cette direction, des représentations du signal vocal récemment suggérées dénommées modèles sinusoïdaux adaptatifs (aSMs) ont été développées. Ce type de représentation adaptative est capable d'ajuster ces paramètres aux caractéristiques locales (phase et / ou amplitude) du signal vocal analysé. Il a été montré que cette famille de modèles sinusoïdaux adaptatifs se révèle robuste et efficace et elle fournit une haute qualité de synthèse vocale. Ce type de représentation a été également étendue à d'autres types d'applications telles que la modification de la prosodie (durée et hauteur) et les résultats ont été très satisfaisants.

Tenant compte des performances des modèles sinusoïdaux adaptatifs, une représentation sinusoïdale adaptative raffinée a été proposée dans cette thèse. Cette représentation est dénommée, représentation adaptative sinusoïdale raffinée (R_aSR). L'étape d'analyse et l'étape d'adaptation de notre nouvelle représentation ont été améliorées. Nous avons utilisé une représentation quasi-harmonique à l'étape de l'analyse pour obtenir une première estimation de ses paramètres instantanés et pour affiner l'estimation de ces paramètres, un schéma adaptatif combiné à un mécanisme itératif de correction de fréquence est utilisé à l'étape d'adaptation. La somme des composantes instantanées estimées donne le signal vocal reconstruit final.

Des tests d'évaluation et des résultats expérimentaux ont confirmé les performances de la représentation suggérée dans la modélisation des signaux vocaux voisés par rapport aux modèles de l'état de l'art. En effet, une comparaison de notre nouvelle représentation du signal de la parole avec les modèles de l'état de l'art stationnaires (modèle SM, modèle HNM) et non stationnaire (modèle adaptatif harmonique-aHM) en matière de rapport signal-erreur de reconstruction (SRER) a été entreprise. Il a été montré que notre représentation et le modèle aHM fonctionnent de manière similaire. Par contre, la représentation (R_aSR) surpasse à la fois le modèle SM et le modèle HNM en matière de SRER. Également et d'un point de vue perceptuel, un test d'écoute a révélé la supériorité de notre représentation par rapport aux représentations stationnaires.

# Perspectives

Dans une grande variété d'applications de la parole, des modifications prosodiques (c'est-à-dire modification de l'échelle de temps et du pitch) sont requises. En effet, de l'industrie cinématographique, du divertissement et des communications, à la synthèse vocale et à la restauration de la voix pathologique, les modifications prosodiques ont reçu une attention croissante et ont été étudiées en profondeur par la communauté du traitement de la parole.

En conséquence, un certain nombre de techniques de modification prosodique ont été proposées dans la littérature, sur la base des modèles correspondants. Ceux-ci appartiennent généralement à deux classes différentes mais non distinctes : les approches paramétriques et les approches non paramétriques. Ces dernières incluent : la technique PSOLA (Pitch-Synchronous Overlap-Add) [73] et ses variantes, telles que la technique WSOLA (Weighted Synchronized Overlap-Add) [74] et MBROLA (Multi-Band Re-synthesis Overlap-Add)[75] ainsi que les techniques à base de vocodeur de phase [76, 77]. Les techniques paramétriques comprennent des modèles à bande étroite, tels que le modèle SM [10], et le modèle HNM [12].

Puisque la modification de la prosodie est d'une grande importance dans le domaine de la synthèse de la parole, et puisque les modèles sinusoïdaux adaptatifs ont été appliqués avec succès à ce type d'application [69, 70], nos prochains travaux de recherche se concentreront alors sur la conception d'un schéma pour la modification du signal de la parole à l'échelle temporelle et fréquentielle, en tenant compte des performances de notre représentation R_aSR dans l'analyse-synthèse vocale .

De plus, puisque les représentations sinusoïdales adaptatives ont été appliquées avec succès dans la modélisation de signaux de parole non voisés, plosives, etc. [64, 65], nous avons pensé aussi à étendre l'utilisation de notre nouvelle représentation R_aSR pour l'analyse-synthèse de ce type de signaux de la parole arabe.

# Bibliographie

# Bibliographie

[1] L. R. Rabiner. Applications of voice processing to telecommunications. Proc. IEE, vol. 82, pp. 199-228, 1994.

[2] S. V. Vaseghi. Multimedia Signal Processing.Theory and Applications in Speech, Music and Communications.Wiley. 2007

[3] J. Mariani. Language and speech processing. ITSE, Wiley. 2009.

[4] T. F. Quatieri. Discrete-Time Speech Signal Processing. Prentice Hall, Engewood Cliffs, NJ. 2002.

[5] T. Dutoit. An Introduction to Text to Speech Synthesis. Kluwer Academic Publishers.1997

[6] P. Taylor. Text-to-Speech Synthesis. Cambridge Univesity Press. 2009.

[7] Y. Tabet, M. Boughazi, S. Afifi. A Tutorial on Speech Synthesis Models. Procedia Computer Science 73. pp. 48-55. 2015.

[8] L. R. Rabiner and R. W. Schafer. Digital Processing of Speech Signals. Prentice Hall, 1978.

[9] G. Fant. Acoustic Theory of Speech Production. Gravenhage,The Netherlands : Mounton and Co. 1960.

[10] R. J. McAulay, T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. IEEE Transactions on Acoustics, Speech, and Signal Processing. vol 34. pp.744-754, 1986.

[11] B. Atal, S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. Journal of Acoustical Society of America (JASA), vol 50. pp.637-655, 1971.

[12] Y. Stylianou. Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[13] Y. Pantazis, O. Rosec, Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In Interspeech, Brisbane, September 2008.

[14] Y. Pantazis, O. Rosec, Y. Stylianou. Adaptive AM-FM signal decomposition with application to speech analysis. IEEE Transactions on Audio, Speech, and Language Processing, vol 19, pp.290-300, 2011.

[15] G. P. Kafentzis, Y. Pantazis, O. Rosec, Y. Stylianou. An Extension of the Adaptive Quasi-Harmonic Model. In Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Kyoto, 2012.

[16] G. Degottex, Y. Stylianou. A full-band adaptive harmonic representation of speech, in Interspeech, Portland, Oregon, U.S.A, 2012.

[17] G. Degottex, Y. Stylianou. Analysis and synthesis of speech using an adaptive full-band harmonic model. IEEE Transactions on Audio, Speech, and Language Processing, vol 21, N°(10) pp.2085-2095, 2013.

[18] G. P. Kafentzis, O. Rosec, Y. Stylianou. Robust full-band adaptive sinusoidal analysis and synthesis of speech. In Proceedings of IEEE International Conference on Acoustic,Speech, and

Signal Processing (ICASSP), 2014.

[19] Y. Tabet, M. Boughazi, S. Afifi . Speech Analysis and Synthesis with a Refined Adaptive Sinusoidal Representation. International Journal of Speech Technology (IJST). 2018.

[20] T. Dutoit. High-Quality Text-to-Speech Synthesis an Overview. Journal of Electrical Electronics Engineering,Australia : Special Issue on Speech Recognition and Synthesis, vol. 17, pp. 25-37, 1999.

[21] Y. Tabet, M. Boughazi . Speech Synthesis Techniques. A Survey. 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), pp. 67-70, 2011.

[22] T .Styger, E. Keller. Formant synthesis. In E. Keller (ed.), Fundamentals of Speech Synthesis and Speech Recognition : Basic Concepts, State of the Art, and Future Challenges (pp. 109-128). Chichester : John Wiley. 1994

[23] D. H. Klatt. Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, vol. 82(3), 1987.

[24] B. Kroger. Minimal Rules for Articulatory Speech Synthesis. Proceedings of EUSIPCO92, pp.331-334,1992.

[25] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. Proceedings of ICASSP-96, Atlanta, vol. 1, pp. 373-376, 1996.

[26] R. Donovan, P. Woodland. Improvements in an HMM-based Speech Synthesizer. Proc. of the EuroSpeech Conf., Madrid, pp. 573-576. 1995.

[27] H. Dudley. The vocoder, Bell Labs. Rec., vol. 17, pp. 122, 1939.

[28] J. Flanagan and Golden. Phase vocoder, Bell Syst. Tech. J., vol. 45, pp. 1493, Nov. 1966.

[29] M. R. Portnoff. Short-time fourier analysis of sampled speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol 29. pp.364-373, 1981.

[30] L. R. Rabiner, R. W. Schafer. Introduction to digital speech processing. Foundations and Trends in Signal Processing. vol. 1, no. 1, pp. 1-194. 2007

[31] P. Hedlin. A tone-oriented voice-excited vocoder. In Proc. IEEE Int. Conf. Accoustic., Speech, Signal Processing. pp. 205-208, Atlanta, 1981

[32] L. B. Almeida, F. M. Silva. Variable-frequency synthesis : an improved harmonic coding scheme. Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), vol 1. pp. 2751-2754, 1984.

[33] R. McAulay, T. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model, in Proc. ICASSP-84 (SanDiego, CA, ). session 27.6.1. Mar. 1984

[34] T. F. Quatieri and R.J. McAuley. Audio signal processing based on sinusoidal analysis/-synthesis. In Mark Kahrs and Karlheinz Brandenburg, editors, Applications of Digital Signal Processing to Audio and Acoustics, chapter 9, pp. 343-416. Kluwer Academic Publishers, 2002.

[35] D. W. Griffin, J. S. Lim. Multiband Excitation Vocoder. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol 36, N°(8), pp.1223-1235, 1988.

[36] A. J. Abrantes, J.S. Marques, IM. Transcoso. Hybrid sinusoidal modeling of speech without voicing decicion. Eurospeeech 91. pp. 231-234, 1991.

[37] W. Oomen, A. C. den Brinker. Sinusoids plus noise modelling for audio signals, In the17th International Conference : High-Quality Audio Coding,August 1999.

[38] J. Laroche, Y. Stylianou, E. Moulines. HNM : A Simple, Effecient Harmonic plus NoiseModel for Speech. InWorkshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA),Pages 169172, New Paltz, NY, USA, Oct 1993.

[39] R. Boite, H. Bourlard, T. Dutoit, J.Hancq, H. Leich. Traitement de la parole. Presses polytechniques et universitaire Romandes, Lausanne, 2000

[40] Caliope. La parole et son traitement automatique. Masson, Paris, 1999.

[41] J. Makhoul. Linear Prediction. A Tutorial Review. Proceedings of the IEEE, vol 63. pp.561-580, 1975.

[42] J. Markel, A. Gray. Linear prediction of speech. Springer Verlag, 1976.

[43] B. S. Atal, J. R. Remde . A New Method of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates, IEEE ICASSP, pp. 614-617. 1982.

[44] B. J Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform, IEEE Trans. on Acoust., Speech, and Signal Proc., vol. ASSP-25, pp. 235-238. 1977.

[45] B. J. Allen, R. L. Rabiner. A Unified approach to short-time fourier analysis and synthesis, Proc. IEEE, vol. 65, pp. 1558-1564. 1977.

[46] M. M. Goodwin. Adaptive Signal Models : Theory, Algorithms, and Audio Applications. PhD thesis. University of California, Berkeley, 1997.

[47] L. R Rabiner and B. Gold. Theory and applications of digital signal processing, Englewood Cliffs, New Jersey : Prentice-Hall,1975.

[48] Oppenheim, V. Alan, R. W. Schafer. Digital signal processing, Englewood Cliffs, New Jersey : Prentice-Hall. 1975.

[49] Dolson, B. Mark. The phase vocoder : A tutorial. Computer Music J., vol. 10, no. 4, pp. 14-27. 1986.

[50] Gerhard Doblinger. Signal processing using MATLAB. Lecture notes, collection of problems, projects. 2014

[51] M. W. Macon. Speech Synthesis Based on Sinusoidal Modeling. PhD thesis, Georgia Institute of Technology, 1996.

[52] M. Crespo, P. Velasco, L. Serrano, J. Sardina. On the Use of a Sinusoidal Model for Speech Synthesis in Text to-Speech. In Progress in Speech Synthesis, pp. 57-70, Springer, 1996.

[53] E. B. George. An analysis-by-synthesis approach to sinusoidal modeling applied to speech and music signal processing, Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1991.

[54] E. B. George, M. J .T. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. IEEE Transactions on Speech and Audio Processing. 5(5). pp. 389-406 , 1997.

[55] X. Serra. A System for Sound Analysis, Transformation, Synthsis based on a Determistic plus Stochastic Decomposition.PhD thesis, Stanford university. 1989.

[56] X. Serra, J. Smith. Spectral modeling synthesis : a sound analysis/synthesis system based on a deterministic plus stochastic decomposition , Computer Music Journal, vol. 14, no. 4, 1990.

[57] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis, IEEE Transactions on Speech and Audio Processing, vol. 9 N°(1), pp.21-29. 2001.

[58] I. Sainz, E. Navas, I. Hernaez. Harmonic plus noise model based vocoder for statistical parametric speech synthesis. Selected Topics in Signal Processing, IEEE. vol.8(2). 2014.

[59] S. Levine. Audio Representations for Data Compression and Compressed Domain Processing. PhD thesis, Stanford University, 1999.

[60] R. Boyer and K. Abed-Meraim, Audio transients modeling by damped and delayed sinusoids (DDS), Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 02), Orlando, Florida, USA, 2002.

[61] H. Thornburg. Detection and Modeling of Transient Audio Signals with Prior Information. PhD thesis, Stanford University, 2005.

[62] Y. Pantazis. Adaptive AM-FM Signal Decomposition With Application to Speech Analysis. PhD thesis, Computer Science Department, University of Crete, 2010.

[63] G. P. Kafentzis. Adaptative sinusoidal models for speech with applications in speech modifications and audio analysis. PhD thesis, Computer Science Department, University of Crete, 2014.

[64] G. P. Kafentzis, O. Rosec, Y. Stylianou. On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi Harmonic Models. In Interspeech, Portland, Oregon,USA, 2013.

[65] G. P. Kafentzis, Y. Stylianou. High-Resolution Sinusoidal Modeling of Unvoiced Speech. In International Conference on acoustics, speech, and signal processing , shanghai, china, 2016.

[66] Y. Pantazis, G. Tzedakis, O.Rosec, Y. Stylianou. Analysis/ Synthesis of Speech based on an daptive Quasi-Harmonic plus Noise Model. In Proc. IEEE ICASSP, Dallas, Texas, USA, 2010.

[67] J. Kominek, A. W Black. The CMU ARCTIC databases for speech synthesis. Tech. Rep. CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003.

[68] N. Halabi. Modern Standard Arabic Phonetics for Speech Synthesis. PhD thesis, University of SOUTHAMPTON, 2016.

[69] G. P. Kafentzis, G. Degottex, O. Rosec, Y. Stylianou. Time-scale Modifications based on an Adaptive Harmonic Model. In Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Vancouver, CA, 2013.

[70] G. P. Kafentzis, G. Degottex, O. Rosec, Y. Stylianou. Pitch modifications of speech based on an adaptive harmonic model. In Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), 2014.

[71] A. Camacho, J. G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. Journal of Acoustical Society of America (JASA), vol124, pp.1628–1652, 2008.

[72] The ITU Radiocommunication Assembly, “Itu-r bs.1284-1 : General methods for the subjective assessment of sound quality,” Tech. Rep., ITU, 2003.

[73] E. Moulines, F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication, vol. 9, pp. 453-467, 1990.

[74] W. Verhelst, M. Roelands. An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech, ICASSP, pp. 554-557, 1993.

[75] T. Dutoit, H. Leich. Improving the td-psola text-to-speech synthesizer with a specially designed mbe re-synthesis of the segments database, in EUSIPCO. pp. 343-347, 1992.

[76] J. Laroche, M. Dolson. Imroved Phase Vocoder Time-Scale Modification of Audio, in IEEE Trans. On Speech and Audio Processing, vol. 7. pp. 323-332, May 1999.

[77] P. Depalle, G. Poirot. SVP : A Modular System for Analysis, Processing and Synthesis of Sound Signals, Proceeding of the 1991 International Computer Music Conference, 1991.