Suivi de Formants par analyse en Multirésolution

Imen Jemaa

Ecole Doctorale IAEM Lorraine

Département de formation doctorale en Informatique — UFR STMIA

Suivi de Formants par analyse en Multirésolution

Thèse

présentée et soutenue publiquement le
pour l'obtention du

Doctorat de l'Université de Lorraine-Nancy I
(Spécialité Informatique)

Par

Jemâa Imen

Composition du jury

Rapporteurs :
CHERIF Adnen, Professeur Université Tunis -El Manar
LIENARD Jean-Sylvain, DR CNRS, LIMSI-CNRS

Examinateurs :
TABBONE Antoine, Professeur, Université de Lorraine
EL AMIRI Hamid, Professeur, Université Tunis El-Manar
LAPRIE Yves, DR CNRS, LORIA
OUNI Kais, Professeur, Université de Carthage
HATON Jean-Paul, Professeur, Université de Lorraine

Table des matières

Liste des figures

Chapitre I

Chapitre II

Chapitre III

Chapitre IV

Chapitre V

Liste des tableaux

Chapitre III

Chapitre V

Introduction générale

Parole et voix sont envisagées de façon pluridisciplinaire, sous des aspects d'ordre physique (traitement du signal, acoustique), linguistique (phonétique, phonologie, syntaxe) ou psychologique (perception et cognition).

Le signal de la parole permet la communication entre humains. Il permet de communiquer la pensée par un système de sons articulés émis par les organes de la phonation et qui est produit par deux processus différents qui sont la vibration des cordes vocales et la turbulence créée par l'air au niveau du conduit vocal (Calliope 1989). L'estimation des caractéristiques du conduit vocal est un domaine de recherche important, notamment à cause de son utilité pour la compréhension et la modélisation du mécanisme de production de la parole.

Lors de la production, c'est le passage de l'air dans le conduit vocal qui va induire la résonance du canal, et générer un groupe de résonances qui sont appelées les formants (Calliope 1989). Les formants sont les paramètres importants qui peuvent être très utiles pour l'identification phonétique, en particulier celles des voyelles et autres sons vocaliques. L'algorithme consistant à repérer et marquer les trajectoires temporelles des fréquences des formants parallèlement à l'axe de temps est appelé suivi de formants. Les formants sont le reflet des caractéristiques individuelles du locuteur. L'estimation des formants et les algorithmes de suivi ont été largement utilisés dans la reconnaissance du locuteur, la synthèse vocale et le codage de la parole (O'Shaughnessy.D 2008). Ce sont donc des sujets de recherche importants dans le domaine du traitement du signal de parole.

Le suivi de formants est un problème qui a vu plusieurs solutions et applications potentielles dans le développement des systèmes de communication basés sur la parole.

En général, le problème du suivi de formants est décomposé en deux étapes. La première étape est dédiée à la tâche d'estimation des fréquences de chaque formant du signal traité. Ces fréquences sont typiquement estimées sous la forme de pics spectraux ou bien à travers les racines du polynôme du modèle de la prédiction linéaire (LPC). L'estimation des pics spectraux est difficile à cause des faux pics et la fusion ou la séparation des formants alors que l'utilisation des racines des modèles prédictifs est limitée par la difficulté de fixer des règles appropriées pour l'identification de la racine considérée comme un formant. Une fois que l'estimation de la localisation temporelle des paramètres formantiques est obtenue, la deuxième étape consiste à relier ensemble les valeurs des fréquences de chaque formant pour regrouper ces fréquences parallèlement à l'axe de temps afin d'obtenir les trajectoires de suivi les plus appropriées pour chaque formant. L'enchaînement des trajectoires peut être accompli en imposant des contraintes de continuité dans (Vargas.J 2008) :

- Les coûts de transition des suivis estimés avec la programmation dynamique (DP) (Talkin.D 1987) (Xia.K 2000)
- Les statistiques d'apprentissage des modèles de Markov cachés (HMM) adaptés (Acero.A 1999) (Toledano.T.D 2006)

L'optimisation des trajectoires basée sur la programmation dynamique est confrontée au problème de la sélection de la puissance propre des contraintes de continuité particulièrement au niveau des frontières des sons. Ce fait a motivé l'inclusion de l'information du contexte phonémique dans l'estimation et le suivi (Lee.M 2005) ; c'est une tâche qui repose sur l'alignement forcé et la segmentation phonémique (Toledano.T.D 2003).

La méthode utilisant les modèles d'appariement des trajectoires de suivi de formants a connu des évolutions marquantes depuis les modèles HMM aux modèles dynamiques cachés HDM (Gang.L 2010). Cette méthode de modélisation de la parole HDM a été proposée par Richards dans (Richards.H.B 1999) pour décrire la structure dynamique du son. Il présente une nouvelle approche de modélisation acoustique-phonétique sous la forme d'un modèle dynamique caché (HDM), qui justifie explicitement le phénomène de coarticulation et les transitions entre les phonèmes voisins. Inspiré par le fait que le discours est réellement produit par un système dynamique sous-jacent, le modèle HDM se compose d'un seul vecteur cible par phonème dans un espace dynamique caché.

Les HDM ont été appliqués avec succès dans la reconnaissance de la parole et de nombreux nouveaux algorithmes de suivi de formants phonétiques ont été élaborés sur cette base (Gang.L 2010)**.** La méthode de Li. Deng (Deng.L 2004) par exemple, a présenté un algorithme de suivi des résonances du conduit vocal (VTR) qui correspondent en fait aux formants surtout pour les voyelles non nasalisées et pour les zones voisées. Par ailleurs les VTR ne fusionnent pas et ne se divisent quelle que soit la partie de discours. Elles correspondent aux fréquences propres du conduit vocal. Les VTR peuvent ne pas correspondre aux proéminences spectrales en présence des zéros au niveau de la fonction de transfert, (par exemple au niveau de certaines fricatives, occlusives et nasales) c'est-à-dire que les proéminences spectrales des VTR peuvent être cachées au niveau du spectrogramme pendant certaines transitions phonétiques. Cependant, les VTR ne disparaissent pas même si le signal de parole ne les renforce pas acoustiquement. Sauf que, la plupart des méthodes de suivi de formants ne prennent en compte que les proéminences spectrales pour faire le suivi.

Pour assurer la continuité du suivi et pour faire une bonne estimation des VTR, Li Deng a suggéré l'utilisation des modèles dynamiques cachés des VTR pour compenser la probabilité d'avoir des fréquences candidates manquantes pendant le suivi. La méthode proposée par Li Deng est basée sur la modélisation d'un modèle de parole structuré basé sur la détection des vecteurs dynamiques cachés des VTR et des paramètres acoustiques qui sont les fréquences et les largeurs de bande correspondantes en utilisant l'analyse LPC. Ensuite, un algorithme itératif de Kalman est mis en œuvre pour effectuer le suivi des VTR et entraîner les paramètres résiduels pour trouver les solutions optimales et améliorer le suivi. Hélas cette approche est limitée par la difficulté de l'estimation des paramètres cibles de modèle HDM.

Dans la littérature, il y a plusieurs autres méthodes qui utilisent le modèle HDM telle que la méthode de Zheng (Zheng.Y 2004). Cette méthode est basée sur un mélange d'états d'un filtre à particules pour le suivi de formants durant les transitions des voyelles et des consonnes. Notamment, l'estimation des formants durant le relâchement des consonnes est très difficile et dans ce cas, une estimation spectrale ARMA n'est pas suffisante et incapable de faire une bonne estimation en présence des zéros. En plus dans la pratique, les fréquences des formants sont cachées par la présence des zéros la plupart du temps pendant la production des consonnes à cause de l'interaction pôle-zéro. A cause de la difficulté d'estimation des zéros spectraux, Zheng a proposé de modéliser les spectres des consonnes et des voyelles par une fonction exponentielle pondérée autorégressive du spectre (EWAR) qui est une fonction capable de présenter avec précision les fréquences et les amplitudes des pôles dans le spectre

ARMA sans la modélisation explicite des zéros. Ensuite, il a proposé le modèle dynamique caché de la production de la parole comme contrainte de continuité dont le but d'estimer les fréquences des formants durant le relâchement des consonnes. Pour assurer un bon suivi des formants, il a proposé un mélange d'états de filtres à particules qui incorpore des connaissances préalables sur dix classes de phonèmes dont le but d'efficacement échantillonner l'espace typique de chaque classe de phonème et par l'utilisation d'une fonction de vraisemblance optimale dans laquelle il a incorporé le modèle dynamique caché afin de réduire le nombre de particules utilisées par le filtre. Cet algorithme est capable de donner une interpolation plausible des fréquences des formants durant le relâchement des consonnes même si la précision du suivi dépend du nombre de particules utilisées.

Il existe d'autres méthodes de suivi de formants qui ont été basées sur les bancs de filtre telle que la méthode de Kumareson (Kumaresan.R 2005) qui est basée sur deux bancs de filtres parallèles inspirés du système auditif. Un ensemble constitue les filtres à bande large qui se chevauchent et l'autre ensemble constitue les filtres à bande étroite. Les deux bancs de filtres coopèrent pour isoler les régions persistantes et transitoires du signal. Les filtres à bande étroite aident à identifier les régions spectrales contenant l'énergie significative du signal et caractériser les composantes transitoires et dans ces régions les groupes des filtres à bande large sont ensuite combinés de façon optimale pour suivre et isoler les résonances des autres fréquences parasites.

Les efforts de ces dernières années sont orientés vers l'estimation des paramètres acoustiques tels que les MFCC. Plusieurs méthodes qui sont actuellement très utilisées et qui ont apporté beaucoup d'améliorations dans ce domaine telle que la méthode de Darch (Darch.J 2005) dans laquelle l'estimation des vecteurs des formants est basée sur les coefficients MFCC et le modèle gaussien GMM (Gaussien Mixture Model). Il y a aussi celle de Bazzi (Bazzi.I 2003) qui est basée sur l'acquisition de la relation entre l'étiquetage des formants et l'information phonétique, qui fait correspondre les paramètres de formants aux coefficients cepstraux MFCC par un prédicteur non linéaire pour établir un dictionnaire (codebook) de prédiction. Pour faire le suivi, cette méthode adopte aussi l'algorithme EM (Maximisation de l'Espérance) pour entraîner les coefficients de l'information résiduelle des signaux de parole et rechercher les paramètres optimaux dans le dictionnaire de prédiction. Cette technique exige un coût de calcul très élevé pour la construction du prédicteur non linéaire.

Dans la littérature, nous remarquons que de nombreux chercheurs se sont appliqués à l'étude de l'acquisition des paramètres des formants et les algorithmes de suivi ces dernières années, et de nouveaux algorithmes sont constamment proposés. Les résultats de la plupart de ces méthodes sont exploités dans les applications du traitement du signal de la parole. Cependant, il existe un manque de bases de données étiquetées de référence en particulier pour langue arabe, qui sont nécessaires à l'évaluation quantitative des techniques automatiques de suivi de formants.

Dans cette étude, nous avons préparé notre base de données en enregistrant un corpus phonétiquement équilibré en langue arabe et en élaborant un étiquetage manuel phonétique et formantique des différents signaux de cette base. Cette dernière est destinée à être utilisée après comme base de référence dans le but est d'évaluer les algorithmes automatiques de suivi de formants testés sur des signaux en langue arabe. Nous avons implémenté une nouvelle approche pour détecter l'ensemble des fréquences formantiques candidates sous formes de crêtes d'ondelette qui sont les maxima du scalogramme en testant trois types d'ondelettes complexes qui sont : Morlet Complexe, Shanon et Frequency B-Spline. Pour réaliser le suivi des trajectoires fréquentielles, nous avons utilisé en premier lieu le calcul du centre de gravité de la combinaison des fréquences formantiques candidates comme première approche de suivi et ensuite, on a utilisé la programmation dynamique combiné avec le filtrage de Kalman comme deuxième approche de suivi. Finalement, nous avons fait une étude exploratoire en utilisant notre corpus étiqueté manuellement comme référence pour évaluer quantitativement nos deux nouvelles approches par rapport à d'autres méthodes de suivi de formants.

Ce rapport s'articule autour de cinq chapitres :

Le premier chapitre sera consacré à une brève description des caractéristiques anatomiques, phonatoires et phonémiques du signal vocal ainsi qu'une classification phonémique pour bien définir les origines et les caractéristiques de chaque phonème de point de vue articulatoire et acoustique. Ensuite, il explorera également l'origine des formants, leurs caractéristiques et l'importance de leur estimation. Finalement, ce chapitre se terminera par la mise en valeur l'intérêt des formants pour la perception de la parole.

Le deuxième chapitre sera consacré à une présentation d'un état de l'art des méthodes d'estimation de formants existantes telles que les méthodes basées sur les approches

paramétriques, sur le lissage cepstral et sur les modèles auditifs. Ensuite, il explorera également un état de l'art des différentes méthodes de suivi de formants existantes.

Le troisième chapitre sera consacré aux différentes étapes pour préparer notre base de données étiquetée en langue Arabe. Il va décrire le corpus et présenter les différentes étapes de l'étiquetage manuel phonétique et formantique ainsi que les difficultés rencontrées lors de cet étiquetage.

Le quatrième chapitre sera consacré à la présentation de notre contribution qui est la méthode de suivi de formants basée sur la détection des crêtes de Fourier qui sont les maxima du spectrogramme, et une autre méthode basée sur la détection des crêtes d'ondelettes qui sont les maxima du scalogramme en testant trois types d'ondelettes complexes tout en utilisant comme contrainte de suivi pour les deux méthodes le calcul de centre de gravité de la combinaison des fréquences formantiques candidates. Finalement, Ce chapitre se terminera par la présentation de notre nouvelle approche de suivi temporel des trajectoires des formants en utilisant la programmation dynamique combinée avec le filtrage de Kalman.

Le cinquième chapitre portera sur la présentation des résultats de notre première approche basée sur les crêtes d'ondelettes en utilisant le calcul de centre de gravité sur des signaux synthétiques en testant trois types d'ondelettes, puis, sur des signaux réels de notre corpus étiqueté. Il mettra aussi en évidence son utilité pour l'évaluation quantitative de nos deux approches de suivi ; celle basée sur le calcul de centre de gravité et l'autre basée sur la programmation dynamique combinée avec le filtrage de Kalman ; comparées à d'autres méthodes automatiques de suivi en prenant les signaux étiquetés issus de la base élaborée, comme référence.

Ce rapport se terminera par une conclusion et des perspectives ainsi que la bibliographie et les annexes utilisés.

Chapitre I
Etat de l'art

1.1. Introduction

Les recherches sur l'analyse de la parole, considérée du point de vue de la perception et du traitement du signal, portent sur la perception du timbre de la voix et des variations de la hauteur tonale, sur l'analyse acoustique de la qualité vocale (signal vocal et effort vocal), et sur les méthodes temps-fréquence de représentation de la parole. Il s'agit de caractériser l'évolution de la fréquence fondamentale (ou pitch) qui évolue en fonction du temps, des formants (résonances du conduit vocal), des durées segmentales, et des paramètres de l'onde glottique lorsqu'on parle plus ou moins fort.

Les formants décrivent les structures spectrales des sons voisés et ils permettent souvent d'identifier le phonème prononcé. Vu leur intérêt, plusieurs techniques ont été proposées pour l'estimation et le suivi des formants et pour faire le lien avec la géométrie du conduit vocal.

Dans ce chapitre nous présentons, une brève description des caractéristiques anatomiques, phonatoires et phonémiques du signal vocal. Ensuite, nous présentons une classification phonémique pour bien définir les origines et les caractéristiques de chaque phonème. Finalement, nous présentons l'origine des formants, leurs caractéristiques, l'importance de leur estimation et leur intérêt pour la perception de la parole.

1.2. Caractéristiques anatomiques phonatoires et phonémiques du signal vocal

Etant donné que le conduit vocal agit comme un filtre acoustique, il donne ainsi au son les indices acoustiques qui distinguent les différents phonèmes. En effet, le signal vocal est représenté par un ensemble de sons qui peuvent être voisés ou non voisés. Par définition, le phonème est la plus petite unité phonique fonctionnelle (Rachedi.J 2005). Les phonèmes sont regroupés par classes selon leur mode et lieu de production, mode de voisement et de nasalité. La réalisation des phonèmes est influencée par le phénomène de coarticulation lié à l'enchaînement d'une suite de sons.

Dans cette section nous décrivons brièvement l'anatomie des organes de la parole, les modes phonatoires, les principales classes phonétiques ainsi que brièvement le phénomène de coarticulation.

1.2.1. Description de l'anatomie des organes de la parole

Pour analyser le signal de parole, il est intéressant de comprendre la façon dont il est produit par le système articulatoire. Le passage de l'air à travers le conduit vocal donne naissance à plusieurs variétés de sons. Les phonèmes diffèrent en fonction de leur lieu et leur mode d'articulation. La Fig.1.1 ci-dessous présente les différents points d'articulation qui sont: les lèvres (labiales), les dents (dentales), les alvéoles (alvéolaires), le palais dur postérieur et antérieur (palatal), la voile du palais (vélaire), la luette (uvulaire), le pharynx (pharyngal), la glotte (glottal), dos (dorsal) et apex (apical). (Calliope 1989) (Dugand.P 1999).

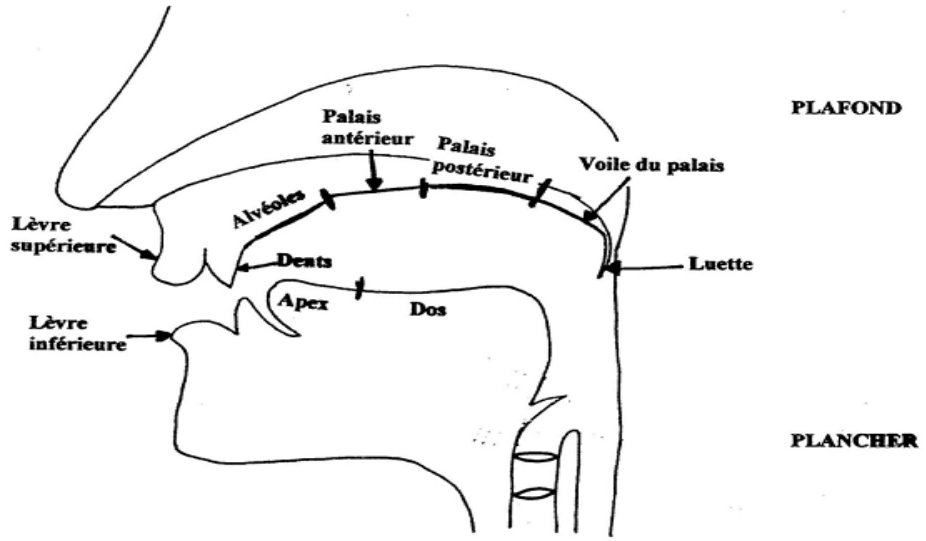


Figure 1. 1 : *Description détaillée de l'appareil vocal*

1.2.2. Les modes phonatoires

« Au cours de la phonation, on peut distinguer différents modes de phonation qui pouvaient être distingués selon l'ajustement laryngé : la position des cartilages aryténoïdes, le degré d'ouverture de la glotte, la tension des cordes vocales et la pression sous-glottique »
(Marchal. A 2007). Les principaux modes phonatoires sont : le voisement, l'aspiration, la voix soufflée, la laryngalisation et l'occlusion glottique.

Le voisement est produit par la vibration des cordes vocales et le rapprochement des cartilages aryténoïdes du larynx (Calliope 1989). Il caractérise les voyelles et certaines consonnes voisées contrairement aux consonnes sourdes qui sont caractérisées par le non voisement. Ce dernier est dû à l'écartement des cordes vocales et l'ouverture totale de la glotte. L'air s'échappe sans être mis en oscillation par les cordes vocales.

L'aspiration est une courte période non voisée (Calliope 1989), caractérisée par une position largement écartée des cartilages aryténoïdes. On peut distinguer des consonnes aspirées et non-aspirées par rapport au délai d'établissement du voisement après le relâchement de l'occlusion ou VOT pour « Voice Onset time » (Marchal.A 2007).

La voix soufflée est produite en écartant les cartilages aryténoïdes l'un de l'autre et l'accolement des cordes vocales est donc incomplet. Une partie de l'air s'échappe sans être modulée et le son voisé est accompagné de souffle : « la voix peut donc manquer d'amplitude » (Marchal.A 2007).

La laryngalisation est un mode phonatoire qui se réalise lorsque les cartilages aryténoïdes sont étroitement rapprochés mais en mettant en vibration seulement une partie des cordes vocales (Calliope 1989). Le son produit est caractérisé par une fréquence fondamentale basse (Marchal. A 2007).

L'occlusion glottique est un mode phonatoire qui se réalise par l'accolement complet des cordes vocales l'une contre l'autre de façon à réaliser une articulation occlusive ce qu'on appelle coup de glotte (Marchal.A 2007).

1.2.3. Classification des sons selon la source : voisés ou non voisés

Selon la nature de la source d'excitation à la sortie du larynx, un signal de parole est tantôt périodique, tantôt aléatoire. Ceci amène à la classification des sons en deux types : voisés ou non voisés (Calliope 1989) (Dutoit.T 2000)**.**

Les sons voisés tels que les voyelles par exemple, sont produits par le passage de l'air des poumons à travers la trachée, qui met en vibration les cordes vocales (Gargouri.D 2010). Pour la parole voisée, l'excitation possède un caractère périodique et des propriétés particulières dues à la forme de l'onde de débit glottique. Les sons voisés sont généralement quasi-périodiques. Ce type de sons représente la majorité du temps de phonation, et est caractérisé en général par une énergie élevée en basse fréquence avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante à l'information linguistique (Dutoit.T 2000).

Par contre, les sons non voisés présentent une structure apériodique, les cordes vocales sont écartées et n'entrent pas en vibration. L'énergie de ce type de sons est concentrée dans les hautes fréquences et correspondant à du bruit (Gargouri.D 2010).

Au niveau du spectrogramme, comme le montre la Fig1.2 ci-dessous, les parties voisées du signal apparaissant sous la forme de successions de pics spectraux denses en énergie sur lesquels on a superposé les courbes des trois premiers formants (F1, F2 et F3), dont les fréquences centrales ne sont pas forcément des multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière.

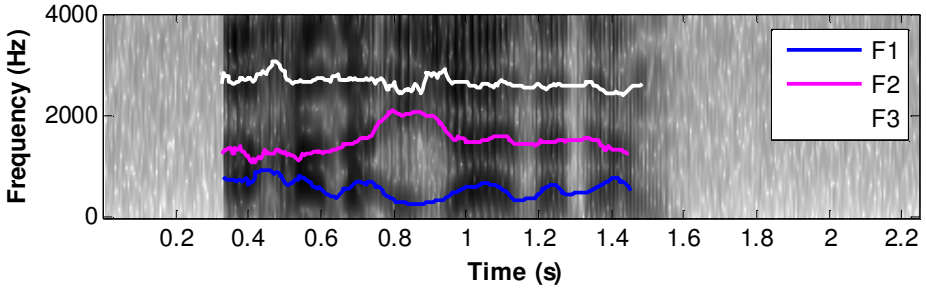


Figure 1. 2 : *Représentation des trois premiers formants (F1, F2 et F3) superposés sur le spectrogramme du signal «* أَخَذَ إِجَازَةً. " 3aχaḍa 3ija:zatan " « Il a pris des vacances» *: où il y a une alternance de sons voisés et non voisés. Dans le cas voisé, une structure formantique est présentée.*

1.2.4. Classification phonémique

« Les langues du monde font un usage spécifique des possibilités physiologiques et anatomiques des organes articulateurs et n'exploitent chacune, qu'une partie des mouvements et positions articulatoires que l'homme peut produire. Ces mouvements articulatoires servent à produire des classes de voyelles et de consonnes distinctes. » (Marchal.A 2007).

La division de l'ensemble de ces sons, ou phonèmes, en classes distinctes, est à l'origine de la constitution d'alphabets phonétiques qui caractérisent les différentes langues. On distingue généralement trois classes principales: les *voyelles*, les *consonnes* et les *semi-voyelles*.

1.2.4.1. Les voyelles

« Si le conduit vocal est suffisamment ouvert pour que l'air expulsé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique » (Dutoit.T 2000)**.** En effet, la voyelle est produite avec un canal aérifère ouvert, sans constriction majeure et en l'absence de génération de bruit de friction. Les voyelles peuvent être spécifiées à l'aide de quatre traits qui sont: la nasalité, le degré d'ouverture du conduit vocal (aperture vocal), la position de la constriction principale du conduit vocal (antérieure/postérieure) et la protrusion des lèvres (arrondissement) (Calliope 1989)**.**

« Les voyelles nasales diffèrent des voyelles orales en ceci que le voile du palais est abaissé pour leur articulation, ce qui met en parallèle les cavités nasale et buccale » (Dutoit.T 2000). On définit et on c1asse les voyelles, comme les consonnes, selon les critères de mode et de point (ou de zone) d'articulation. « Trois positions extrêmes serviront à définir les

voyelles principales. En schématisant, elles occupent les sommets d'un triangle qu'on appelle le « triangle vocalique » (Dugand.P 1999). (Voir Fig1.3)

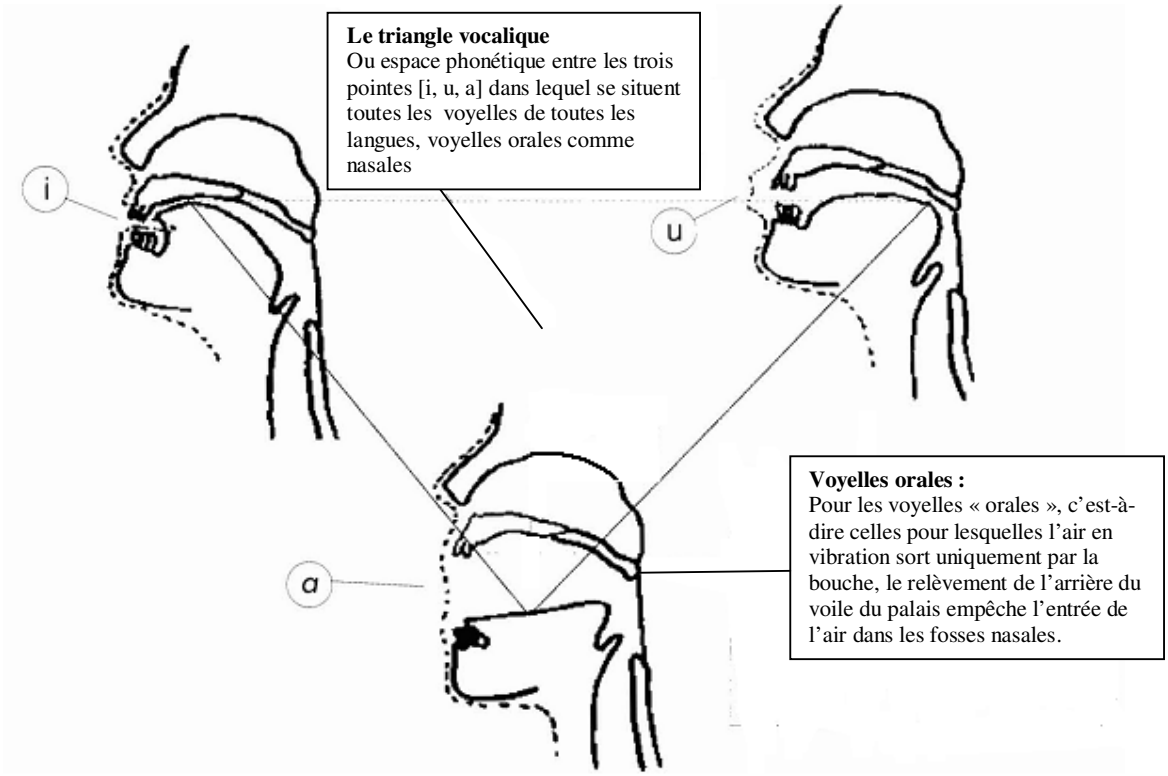


Figure 1. 3 : *Présentation des trois voyelles principales dans le triangle vocalique* (Knoerr.H 2011)

[i] : voyelle orale fermée antérieure rétractée.

[a] : voyelle orale ouverte

[u] : voyelle orale fermée postérieure arrondie

La phonétique classique classe les voyelles d'après la position de la langue dans la cavité buccale. Pour les caractériser, elle prend également en compte la position et la forme des lèvres. Ces derniers peuvent être plus au moins écartées, étirées, arrondies et protruites à des degrés divers. En effet les lèvres peuvent (Knoerr.H 2011):

-« soit se projeter en avant et s'arrondir, et la voyelle est une voyelle arrondie ou labialisée, comme /y, œ, o, u/ ».

- « soit s'étirer ou rester en position neutre : la voyelle est alors une voyelle non labialisée ou non arrondie, comme /i, e, ɛ, a/ ».

Il faut noter qu'il existe une certaine corrélation entre la hauteur de la langue et la labilaité (Marchal.A 2007).

On distingue aussi, selon les mouvements horizontaux de la langue dans la bouche, les voyelles antérieures (en avant), les voyelles centrales (au milieu), et les voyelles postérieures (en arrière), et, selon l'écartement entre la langue et le lieu d'articulation appelé aperture (on parle donc de degré d'aperture des voyelles), les voyelles fermées et ouvertes. Toutes les voyelles de toutes les langues sont toujours situées dans le triangle vocalique, orales comme nasales. (Voir Fig1.4)

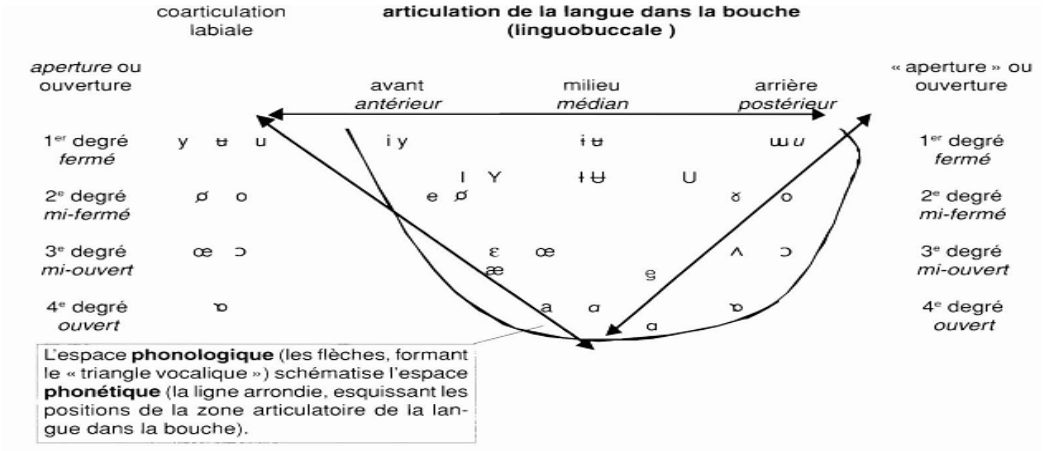


Figure 1. 4 : *Triangle vocalique : Les voyelles selon l'alphabet phonétique international (API)* (Knoerr.H 2011)

Le triangle vocalique constitue un moyen de repérage facile pour placer les voyelles. On distinguera des niveaux d'ouverture (de 2 à plus de 5) et des niveaux de profondeur (de 2 à 7: 4 en Français). Notons que l'indication antérieure / postérieure n'a pas de sens pour la voyelle ouverte étant donné que la langue est abaissée et est alors centrée par rapport à la profondeur (Dugand.P 1999)**.**

L'aperture vocalique désigne la distance verticale qui sépare le sommet du dôme de la langue et le palais. Elle renvoie au degré de courbure convexe de la surface de la langue et à sa hauteur relative dans la cavité buccale. Sur cette base on peut distinguer des voyelles hautes comme /i/ et /u/, mi-hautes comme /e/ et /o/, mi-basses comme /ɛ/ et /ɔ/ et basses comme /a/ et /ɑ/. Il s'agit donc d'un classement qui repose sur une dimension verticale et non sur l'aire au lieu d'articulation comme pour les consonnes (Marchal.A 2007)**.**

La plupart des voyelles utilisées dans les langues sont sonores, c'est-à-dire qu'elles sont prononcées en faisant vibrer les cordes vocales, mais des voyelles sourdes, sans vibration des cordes vocales, sont utilisées dans certaines langues comme le Cheyenne et le japonais. Le chuchotement utilise aussi, par définition, des voyelles sourdes.

La voyelle constitue le pivot d'une syllabe. Tandis qu'une voyelle peut constituer à elle seule une syllabe cela ne peut être le cas d'une consonne. Celle-ci doit être nécessairement associée à une voyelle.

1.2.4.2. Les consonnes

Comme leur nom l'indique « consonna : 'dont le son se joint avec ' », elles précèdent ou suivent un élément vocalique (Marchal.A 2007). Lorsque le conduit vocal au passage de l'air se rétrécit par endroits, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, de leur voisement et leur nasalisation (Calliope 1989). Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général: l'organe articulatoire, la source sonore, l'intensité, l'aspiration, la palatalisation, et la direction du mouvement de l'air (Dutoit.T 2000). Les consonnes forment donc une classe très hétérogène que l'on peut décomposer en trois sous-classes principales ayant des caractéristiques distinctes: *les fricatives*, *les occlusives* et *les sonantes* (Calliope 1989).

- Les consonnes **fricatives** ou constrictives prennent naissance dans la production d'un bruit de friction qui résulte d'une turbulence aérodynamique en un ou plusieurs points du conduit vocal en raison de la présence d'un fort resserrement (ou constriction) dans le flot d'air expiratoire (Calliope 1989). « Ce sont essentiellement les lèvres et la langue qui, selon leur position et leur tension musculaire particulière, conditionnent le type de friction réalisée » (Linguistique UNIL 2011). Cette friction est réalisée au niveau d'un lieu d'articulation qui peut être le palais [ʃ, 3], les dents [s, z], ou les lèvres [f, v], elle produit un bruit en haute fréquence. On parle généralement de fricatives non voisées et des fricatives voisées. Les fricatives non voisées sont caractérisées par le passage d'un écoulement d'air à travers la glotte ouverte, tandis que dans le cas des fricatives voisées, la source vocale est active (Calliope 1989), il y aura donc une vibration des cordes vocales incomplète, c'est-à-dire que les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais

complète et combinée avec un bruit de friction : c'est une combinaison de composantes d'excitation périodique et turbulente (Dutoit.T 2000).

- Les consonnes **occlusives** se caractérisent principalement par un silence dû à la fermeture complète du conduit vocal appelée « occlusion » en un lieu bien défini. (Calliope 1989). Une forte pression est créée en amont de cette occlusion qui peut être au niveau du palais [k, g], des dents [t, d], ou des lèvres [p, b]), puis relâchée brusquement (Dutoit.T 2000). La période d'occlusion est appelée la phase de tenue. Les occlusives peuvent être généralement soit des occlusives voisées (sonores) ou non voisées (sourdes). Les occlusives voisées sont en général plus brèves au niveau du silence que les occlusives sourdes (Calliope 1989). Dans le cas des occlusives voisées par exemple [b, d, g], le silence n'est pas total dans la mesure où la vibration des cordes vocales pendant la tenue articulatoire se traduit par la « barre de voisement » qui est une faible énergie en très basse fréquence au niveau du spectrogramme. Les consonnes occlusives non voisées, par exemple [t, k], sont caractérisées par un silence pendant la tenue articulatoire suivi par une explosion au moment de relâchement (Dutoit.T 2000). « Conventionnellement, on mesure « le délai d'établissement du voisement » (VOT : Voice Onset Time) à partir de la barre d'explosion » (Calliope 1989). « La mesure est comptée négativement si le voisement précède la barre d'explosion, et positivement dans le cas contraire ». Les consonnes voisées par exemple [b, d, g] sont à VOT négatif et les consonnes non voisées [t, k] sont à VOT positif. Les transitions formantiques des occlusives sont caractérisées par des déflexions fréquentielles rapides des formants que l'on observe au passage d'une consonne à une voyelle et réciproquement et cela est dû à la diminution du degré de constriction qui suit la rupture de l'occlusion et les mouvements des organes articulatoires vers une nouvelle cible (Calliope 1989).

- Les consonnes **sonantes** (Calliope 1989) possèdent comme les voyelles, une structure formantique, mais au contact des consonnes sourdes, les sonantes qui sont intersèquement sonores perdent leur voisement. Cette classe est en fait constituée du regroupement de deux sous-classes que sont les liquides et les nasales :

  -Les liquides [l, r] sont assez difficiles à classer. L'articulation de [l], qui une latérale, ressemble à celle d'une voyelle, mais la position de la langue conduit à une fermeture partielle du conduit vocal. Le son de [r], quant à lui, admet plusieurs réalisations fort

différentes qui peuvent être soit trille c'est-à-dire répétitive au niveau du spectrogramme, soit tap avec la tenue d'un silence de durée très limitée. Généralement, au niveau du spectrogramme, on observe des formants vocaliques pour les liquides.

-« Les nasales [m, n] font intervenir les cavités nasales par abaissement du voile du palais. » (Dutoit.T 2000) « Les spectrogrammes de nasales présentent de nombreuses similitudes avec ceux des voyelles; on y découvre, le plus souvent, un premier formant fort en basse fréquence » (Linguistique UNIL 2011).

1.2.4.3. Les semi-voyelles

Les semi-voyelles [j, µ, w] combinent certaines caractéristiques des voyelles et des consonnes, on les appelle parfois semi-consonnes ou glides (Dugand.P 1999). D'ailleurs on peut les classer selon la sous-classe sonantes des consonnes. « Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes » (Dutoit.T 2000). Les fréquences initiales des formants d'une semi-voyelle sont proches de celles de la voyelle correspondante.

1.2.5. Phénomène de coarticulation

« La parole est produite par une chaîne de gestes articulatoires qui se réalisent dans le temps. » (Marchal.A 2007). Cependant, les sons de la parole sont caractérisés par une importante variabilité selon leur entourage phonétique. (Nguyen.N 2001) Cette variabilité a été en partie attribuée au fait que les mouvements accomplis par les articulateurs dans la production de la parole se chevauchent sur l'axe temporel, par exemple, en prononçant la syllabe CV, les gestes articulatoires associés à la consonne initiale et à la voyelle qui la suit sont partiellement superposés. « Ces phénomènes de recouvrement temporel sont généralement désignés sous le terme de coarticulation ». (Nguyen.N 2001) En effet un phonème ne se prononce pas de la même manière, tout dépend s'il est prononcé seul ou dans une chaîne parlée (nommé « allophone ») et tout dépend de son entourage phonétique. « Cependant, la modification de la configuration du conduit vocal pour passer d'un phonème à un autre se fait de façon progressive et les deux sons subissent une distorsion » (Marchal.A 2007). « Il faut noter aussi que les articulations se succèdent très rapidement : une première articulation peut ne pas être achevée au commencement de la seconde et la représentation de

la réalisation de chaque phonème peut varier considérablement en fonction de son voisinage » (Linguistique UNIL 2011).

Il est reconnu que les phonèmes s'influencent aussi bien à droite par rétention de l'articulation qu'à gauche par anticipation. Par exemple, une consonne sonore suivie d'une consonne sourde a tendance à se dévoiser. Ainsi le mot « médecin », après la chute du « e » devient « médcin » où le « d » perd sa marque de sonorité (Marchal.A 2007); d'où l'effet du phénomène de coarticulation. « Le phénomène de coarticulation est primordial à l'intelligibilité de la phrase. En effet, si on synthétise une phrase en mettant bout à bout les phonèmes composant cette phrase, mais sans aucune contrainte de coarticulation entre phonèmes, on obtient une phrase devient incompréhensible » (Marchal.A 2007).

Les phénomènes de coarticulation ont donné lieu à de nombreuses études dans le domaine de la perception. On sait ainsi que la perception des fricatives par exemple, est soumise à l'influence du degré d'arrondissement labial de la voyelle suivante. (Nguyen.N 2001) Les effets de coarticulation constituent une source d'information mise à profit dans le traitement de la parole. Ils sont suffisamment marqués pour donner à un auditeur la possibilité d'identifier correctement des syllabes dont une portion a été supprimée. (Nguyen.N 2001)

En revanche, à cause du phénomène de coarticulation, il est souvent difficile de segmenter le signal de parole en unités discrètes pour l'analyse, ce qui serait très utile dans les applications automatiques de reconnaissance de la parole (O'Shaughnessy.D 2008). Dans certains cas, une telle division est facile, surtout lorsque l'excitation change brusquement avec le début ou la fin de la vibration des cordes vocales (c'est-à-dire une transition voisée-non voisée), ou au moment de l'ouverture et fermeture du conduit vocal (par exemple, la fermeture des lèvres). (O'Shaughnessy.D 2008).

1.3. Caractéristiques fréquentielles du signal de parole

L'analyse dans le domaine fréquentiel désigne l'analyse des fonctions mathématiques ou des signaux selon la fréquence, plutôt qu'une fonction de temps. Une fonction donnée ou un signal peuvent être convertis entre les domaines temporels et fréquentiels à l'aide d'un opérateur mathématique appelé une transformation. Un exemple est la transformée de Fourier, qui décompose une fonction en la somme d'un nombre (potentiellement infini) d'ondes sinusoïdales. Le «spectre» est la représentation dans le domaine de fréquentiel du signal. La transformée de Fourier inverse convertit la fonction du domaine fréquentiel à une fonction

dans le domaine temporel. Un analyseur de spectre est l'outil couramment utilisé pour visualiser le monde réel des signaux dans le domaine fréquentiel.

On affiche généralement la façon dont le spectre de parole change au fil du temps sous la forme d'un spectrogramme comme le montre la Fig.1.5. Le but de l'estimation spectrale est de décrire la distribution fréquentielle de la puissance contenue dans un signal.

Le signal de parole est caractérisée comme suit (Anusuya.M.A 2011)**:**

• La bande passante du signal est d'environ 10 kHz.

• Le signal voisé est périodique avec une fréquence fondamentale entre 80 Hz et 350 Hz.

• La distribution spectrale présente des pics d'énergie

• L'enveloppe du spectre de puissance du signal montre une atténuation quand la fréquence s'élève (-6 dB par octave).

1.3.1. Bande passante

La bande passante du signal de parole est beaucoup plus élevée que 4 kHz. Pour les fricatives, il ya une quantité importante d'énergie en haute fréquence. Une bande passante de 4 kHz qui est celle du téléphone, contient toutes les informations nécessaires pour comprendre la voix humaine sauf certaines fricatives.

1.3.2. Fréquence fondamentale

Le cycle de vibration des cordes vocales se compose grossièrement d'une période de fermeture, puis d'une période d'ouverture de la glotte. Le phénomène vibratoire est auto-entretenu, ce qui confère au système un fonctionnement pseudopériodique et non strictement périodique. Une étude détaillée des mécanismes productifs, physiologiques et aérodynamiques, mis en œuvre est exposée dans (Calliope 1989).

La fréquence fondamentale (souvent appelée « pitch ») est l'inverse de la période laryngienne, c'est-à-dire du temps séparant deux instants consécutifs de fermeture de la glotte et elle se traduit par un train d'impulsions. La fréquence fondamentale du signal de parole est notée F0. La mesure de F0 n'est évidemment possible que pour les sons voisés.

1.3.3. Fréquence des formants

Les cordes vocales produisent l'énergie sonore qui est résonnée et filtrée par des cavités du conduit vocal : les cavités nasale, buccale, labiale et pharyngale. On notera que certains résonateurs ont une géométrie variable (Dugand.P 1999). Chaque résonateur possède

sa propre fréquence propre. Les fréquences amplifiées sont celles qui sont voisines de la fréquence propre de la cavité. Ce sont les fréquences renforcées que l'on nomme formants. (Voir Fig1.5).

Un formant est un renforcement spectral créé par une cavité du conduit vocal (telle que la bouche, le pharynx…) agissant comme un résonateur. Les formants décrivent les structures spectrales des sons voisés qui nous permettent d'identifier le type du son prononcé en particulier pour les voyelles (Acero.A 1999) ou les autres sons vocaliques (Ali.JA.M.A 2002). En outre, il est bien établi que les formants sont liés à la forme géométrique du conduit vocal.

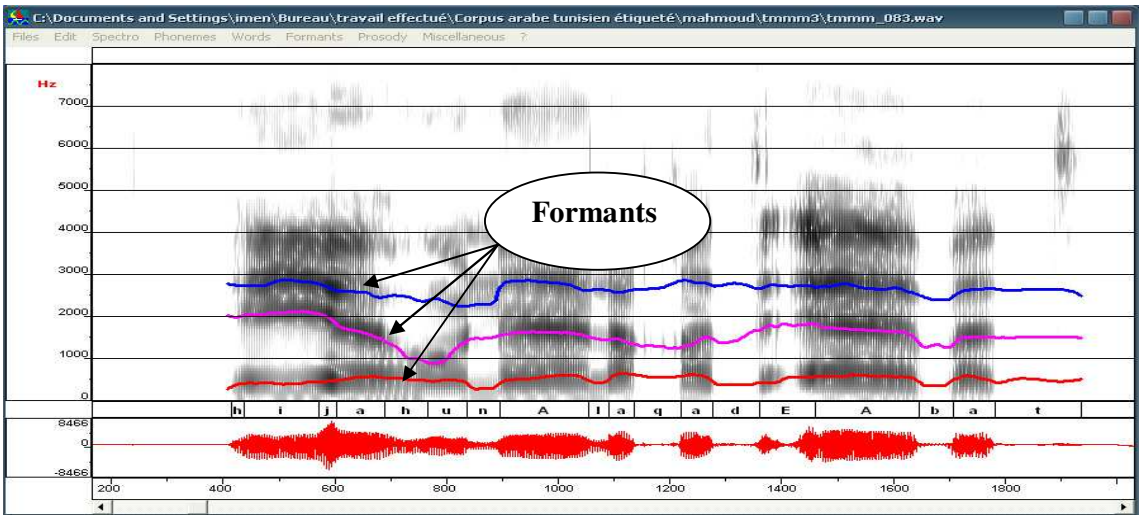


Figure 1. 5 : *Représentation des trois premiers formants superposés sur le spectrogramme du signal «* هِيَ هُنَا لَقَدْ آبَتْ "hiya huna: laqad 3a:bat" « Elle est ici et elle est pieuse »

1.3.4. Variabilités des fréquences des Formants

Dans cette section, les origines et les caractéristiques des formants sont explorées en fonction de leur comportement dans les différents types de sons (Rekhis.O 2009).

Le signal de parole produit est non stationnaire c'est-à-dire variable temporellement changeant de caractéristiques chaque fois que les muscles du conduit vocal se contractent ou se relâchent (Gargouri.D 2010). Pour la plupart des sons, le conduit vocal est constitué des cavités pharyngienne et buccale qui sont les deux principaux résonateurs, plusieurs formants sont donc créés dans le timbre résonantiel, ce qui est d'ailleurs la condition nécessaire à la constitution des voyelles. Lorsque le voile du palais est abaissé, le conduit nasal vient se brancher en dérivation avec le conduit oral après le pharynx, permettant la génération des voyelles nasales et des consonnes nasales introduisant des anti-formants (Léothaud.G 2005). En effet, le conduit vocal peut être modélisé comme un tube acoustique avec des résonances

appelées formants et des creux ou vallées appelés anti-formants.

On peut distinguer plusieurs sources de variabilité des fréquences formantiques, liées à des différences physiologiques entre les locuteurs selon le genre et l'âge, aux effets de la coarticulation et donc des consonnes voisines à la voyelle, et à l'effet émotionnel ou environnemental.

En effet, la parole est principalement produite grâce aux cordes vocales qui génèrent la fréquence fondamentale (le pitch). Cette fréquence de base sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre. Par exemple la voix d'un homme est plus grave que celle d'une femme, ce qui traduit le fait que la fréquence fondamentale est plus faible (Gargouri.D 2010)**.** Le conduit vocal présente aussi d'autres différences, il peut être de forme et de longueur variables selon les phonèmes et les individus et plus généralement selon le genre et l'âge. Ainsi le conduit vocal féminin est en moyenne 15% plus court qu'un conduit masculin typique (Calliope 1989) et donc pour une femme les formants devraient être de 15% plus élevés que ceux des hommes (Gargouri.D 2010). La fréquence et le nombre des formants dépendent de la longueur du conduit vocal. La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

La variabilité intra-locuteur concerne les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut également dégrader la qualité du signal vocal. L'humeur ou l'émotion du locuteur peuvent aussi influencer son rythme d'élocution, son intonation ou sa phraséologie et par suite influe sur les fréquences des formants estimés. La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution. Cette variabilité est considérée comme du bruit. Il existe un autre type de variabilité intra-locuteur lié à la phase de production de parole. Cette variation est due au phénomène de coarticulation.

1.3.5. L'intérêt des formants pour la perception des sons de la parole

Chaque son se distingue par sa structure propre et unique dans le domaine spectral. Pour les phonèmes voisés, la signature implique de fortes concentrations en énergie appelée formants**.**

Les caractéristiques des formants sont spécifiques pour chaque son, d'où leur intérêt pour la perception et la classification des différents sons.

Les formants d'une voyelle composent le patron formantique qui est essentiel pour la reconnaissance des voyelles. En comparant les deux premiers formants F1 et F2 (Calliope 1989) on peut déterminer la nature de la voyelle émise; une augmentation de F1 correspond à une ouverture articulatoire et une augmentation de F2 présente une antériorisation de l'articulation (elle détermine aussi les lieux d'articulation des consonnes). Dans les voyelles, F1 peut varier de 300 Hz à 1000 Hz. La voyelle /i/ par exemple du mot anglais «beet», a une valeur de F1 basse vers 300 Hz. F2 peut varier de 850 Hz à 2500 Hz, la valeur F2 est grossièrement proportionnelle à l'antériorité ou la postériorité de la partie haute de la langue pendant la production de la voyelle. En outre l'arrondissement des lèvres, provoque un F2 plus bas qu'avec des lèvres non arrondies. Par exemple, /i/ comme dans «beet», le mot a un F2 de 2200 Hz, ce qui est le maximum pour F2. Dans la production de cette voyelle, la pointe de la langue est assez loin en avant et les lèvres sont non arrondies. À l'extrême opposé, /u/ comme dans le mot « boot » a un F2 de 850 Hz pour cette voyelle, la pointe de la langue est très loin en arrière, et les lèvres sont arrondies (Anusuya.M.A 2011). La corrélation entre facteurs articulatoires et facteurs acoustiques signifie que le timbre des voyelles est du essentiellement aux deux premiers formants, Il est donc commode de représenter les voyelles dans le plan F1-F2 (Dutoit.T 2000) (Voir Fig.1.6):

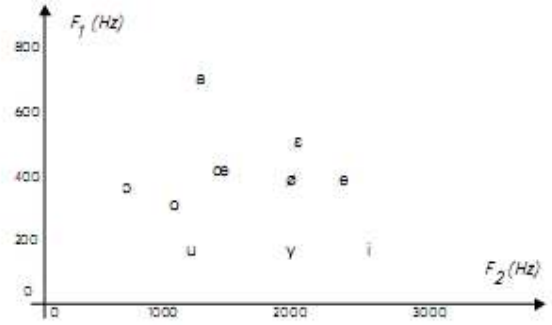


Figure 1. 6 : *Représentation des voyelles dans le plan F1-F2.*

Par ailleurs, des expériences de synthèse ont montré que la position fréquentielle des trois premiers formants caractérise aussi le timbre vocalique (Calliope 1989); pour les voyelles antérieures on observe une baisse de F2 et F3 qui indique la configuration des lèvres. Les formants supérieurs tels que F4 et F5 ont un rôle significatif dans la détermination de la qualité vocale. Il est à noter que les sons nasalisés sont voisés et que certaines voyelles peuvent être partiellement nasalisées. Cependant, la modélisation acoustique des voyelles nasales doit tenir compte de la présence des cavités qui introduisent plusieurs pôles-zéros supplémentaires. La nasalisation est une caractéristique importante pour l'identification des

voyelles qui correspond notamment à un abaissement du formant F1.

L'identification des consonnes est un problème plus complexe que l'identification des voyelles. Acoustiquement, les consonnes sonantes se caractérisent par une structure de formants comme pour nasales (/m/ et /n/) et les latérales (/l/)**.** Pour les nasales par exemple, grâce à la fréquence de résonance basse de la cavité nasale, l'énergie est généralement localisée en basse fréquence. Cette zone représente généralement le premier formant F1.
Il est clair de ce qui précède que les fréquences des formants jouent un rôle majeur dans l'identification des sons de la parole et surtout des voyelles. Cependant, pendant la production de la parole continue, il existe des transitions rapides d'un son au suivant, il est donc important de considérer les effets de ces transitions d'où l'importance du suivi de formants.

1.4. Conclusion

Nous avons présenté dans ce chapitre, une brève description des caractéristiques anatomiques, phonatoires et phonémiques du signal vocal. Ensuite, nous avons présenté une classification phonémique pour bien définir les origines et les caractéristiques de chaque phonème du point de vue articulatoire et acoustique. Finalement, nous avons terminé par explorer l'origine des formants, leurs caractéristiques, l'importance de leur estimation et leur intérêt pour la perception de la parole.

Chapitre II
Méthodes existantes d'estimation et de suivi des formants

2.1. Introduction

Un formant est un renforcement spectral créé par une cavité du conduit vocal (telle que la bouche, le pharynx…) agissant comme un résonateur. Les formants décrivent les structures spectrales qui nous permettront d'identifier le type du son prononcé en particulier pour les voyelles (Acero.A 1999) et autres sons vocaliques (Ali.JA.M.A 2002)**.** En outre, il est bien établi que les formants sont liés à la forme géométrique du conduit vocal. Afin d'améliorer les modèles de production et d'étudier l'identification des sons, il est important d'avoir une estimation précise des formants et de leurs transitions (Plante.F 1994). Le suivi de formants vise à trouver les trajectoires formantiques au cours du temps. Il joue un rôle très important pour le traitement du signal de parole; les formants permettent en effet de représenter la parole de manière concise et utile pour le codage de la parole, et de définir des indices forts pour la classification des phonèmes. La représentation formantique a été utilisée essentiellement en synthèse et en modification de la parole (Alessandro.C 1992) et ce sont également des paramètres pour la reconnaissance.

Etant donné l'intérêt potentiel des formants, de nombreux efforts ont été consacrés au développement d'algorithmes de suivi de formants. L'estimation automatique des formants reste délicate malgré l'emploi de diverses techniques telles que les techniques paramétriques

basées sur la prédiction linéaire et les vecteurs MFCC, et d'autres techniques non paramétriques basées sur le lissage cepstral et les modèles auditifs. Pour avoir un algorithme de suivi de formants performant, il faut une estimation précise des formants à chaque instant et un bon algorithme pour reconstruire des trajectoires grossièrement parallèles à l'axe du temps. Plusieurs algorithmes de suivi sont basés sur les approches probabilistes et la segmentation phonémique. Ils feront l'objet de la section suivante.

Dans ce chapitre nous présentons tout d'abord, des techniques d'estimation des fréquences des formants basées principalement sur l'analyse de la prédiction linéaire, le lissage cepstral et les modèles auditifs. Ensuite, nous présentons certaines techniques de suivi de formants telles que celles basées sur la programmation dynamique, d'autres basées sur les modèles HMM et on terminera par présenter quelques approches de suivi basées sur le filtrage de Kalman.

2.2. Les techniques d'estimation des formants

Dans la bibliographie, plusieurs approches ont été développées pour estimer les formants. Généralement, on estime seulement les trois premiers formants, parce qu'ils décrivent l'information la plus utile dans un signal de parole.
Dans la littérature il existe plusieurs méthodes d'estimation des formants dont les plus utilisées sont :

- Les algorithmes basés sur l'analyse de la prédiction linéaire LPC,
- Les algorithmes basés sur l'analyse cepstrale,
- Les algorithmes basés sur les modèles auditifs.

2.2.1. Estimation des formants par prédiction linéaire

Le signal vocal produit par le système phonatoire présente des caractéristiques propres qu'il est utile d'exploiter lors de l'analyse. Il semble ainsi intéressant de concevoir des méthodes d'analyse spécifiques fondées, d'une manière ou d'une autre sur le processus de production de la parole. Le cadre théorique de la prédiction linéaire est celui d'une

modélisation des relations entre les paramètres articulatoires et acoustiques des sons de parole à l'aide d'une approximation linéaire du processus de production de la parole.

2.2.1.1 Principe du modèle de prédiction linéaire

Il est commode de simplifier le modèle acoustique linéaire de production en réunissant dans un même filtre les contributions de la glotte, du conduit vocal, du rayonnement, et en réduisant l'excitation à un train périodique d'impulsions pour la parole voisée (Alessandro.C 1992).

$$S(z) = E(z)H(z) \qquad \text{(Eq.2.1)}$$

Dans ce modèle simplifié (voir l'équation.2.1), la propriété globale (train périodique d'impulsions ou bruit blanc) de l'excitation est comprise dans la source de spectre plat *E*, alors que les propriétés particulières de cette excitation, se joignent à l'action du conduit vocal et du rayonnement dans le filtre *H*. Ce filtre linéaire évolue dans le temps. Les organes articulateurs (velum, langue, lèvres, mâchoires) provoquent des évolutions rapides de la conformation du conduit vocal, et donc de ses propriétés acoustiques. La vitesse d'évolution des fréquences formantiques peut atteindre 1 kHz en 60 ms (16,7 kHz/seconde).

L'objectif des représentations spectrales qui font référence aux modèles de production de la parole est essentiellement d'identifier les paramètres liés aux sources d'excitation et au filtre linéaire évoluant dans le temps associé au conduit vocal. Les modèles qui vont suivre seront examinés à la lumière de cette décomposition (Alessandro.C 1992).

La prédiction linéaire (LPC : Linear Predictive Coding) (Delsuc.M.A) (Haton.J.P 2006) du signal est une technique largement utilisée en traitement de la parole. Elle se fonde sur la corrélation entre les échantillons successifs du signal vocal. En effet, l'échantillon *s(n)* à l'instant *n* peut être prédit approximativement comme une combinaison linéaire des *p* d'échantillons précédents :

$$s(n) \approx \hat{s}(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots a_p s(n-p) = \sum_{i=1}^{p} a_i s(n-i) \qquad \text{(Eq.2.2)}$$

où a_i représente les coefficients de prédiction et *p* l'ordre de prédiction.

Les coefficients de prédiction a_i sont supposés constants sur une fenêtre d'analyse du signal. En introduisant une excitation normalisée $v(n)$ et un gain d'excitation *G* on obtient :

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gv(n) \qquad \text{(Eq.2.3)}$$

Où $Gv(n)$ est identifiée à l'erreur de prédiction introduite par le modèle ou résidu d'ordre *p* :

$$e(n) = s(n) - \hat{s}(n) \qquad \text{(Eq.2.4)}$$

Soit en transformée en z :

$$S(z) = \left(\sum_{i=1}^{p} a_i z^{-i} \right) S(z) + GV(z) \qquad \text{(Eq.2.5)}$$

Où $V(z)$ désigne la transformée en z de $v(n)$

Une simplification supplémentaire du modèle linéaire de production (2.1) amène à définir un filtre linéaire tout-pôle dont la fonction de transfert est $H = 1/A$:

$$H(z) = \frac{S(z)}{E(z)} = \frac{S(z)}{GV(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad \text{(Eq.2.6)}$$

Le filtre obtenu par ce modèle linéaire simplifié de production est équivalent à un filtre prédictif, ou modèle autorégressif tout-pôle du signal (Dutoit.T 2000). Ce modèle peut être assimilé au modèle acoustique linéaire de production de parole. (Voir Fig.2.1 ci-dessous). Ce modèle est formé de la concaténation de tuyaux sonores de sections différentes et variables selon les sons à produire. La fonction d'excitation $v(n)$ est soit un train d'impulsions quasi périodiques produites par la vibration des cordes vocales pour les sons voisés, soit une source de bruit aléatoire pour les sons non voisés.

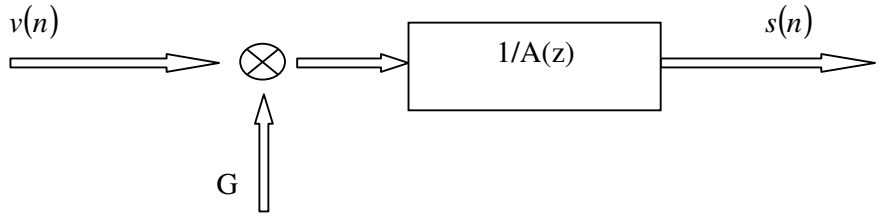


Figure 2. 1 : *Le modèle autorégressif de prédiction linéaire de parole*

On s'intéresse à l'erreur quadratique moyenne commise sur un ensemble de *n* échantillons de parole autour de $s(n)$:

$$E_n = \sum_m \left[s_n(m) - \sum_{i=1}^{p} a_i s_n(m-i) \right]^2 \qquad \text{(Eq.2.7)}$$

Pour minimiser E_n, on annule ses dérivées partielles par rapport aux coefficients de prédiction a_i ($a_0 = 1$) ce qui fournit un système de *p* équations à *p* inconnues dits de Yule-Walker. La résolution de ce système d'équations peut être effectuée à l'aide de diverses méthodes rapides qui relèvent de deux familles :

-La méthode de covariance qui suppose le signal *s* connu seulement sur une plage temporelle de durée limitée.

-La méthode d'autocorrélation dans laquelle la plage d'existence de *s* est infinie. Pour résoudre le système, cette méthode préserve le caractère Toeplitz de la matrice R, en prenant un estimateur ergodique (Pinquier.J 2004). Notons que la résolution des équations de Yule-Walker permet de garantir le minimum de phase au filtre de prédiction linéaire (tous les pôles de filtre sont à l'intérieur du cercle unité), ce qui entraîne la stabilité du filtre. L'ensemble des équations peut être résolu plus efficacement en utilisant des méthodes qui tiennent profit de la structure symétrique de la matrice d'autocorrélation, telles que Levinson ou Levinson-Durbin. L'algorithme de Levinson est une version rapide de la résolution des équations en effectuant une récurrence sur l'ordre du modèle (Jemâa.I 2007).

Les deux méthodes donnent des résultats comparables pour des fenêtres d'analyse suffisamment longues (quelques dizaines de ms). En revanche, la méthode de covariance est préférable pour l'analyse des sons brefs (par exemple une période de voisement) (Haton.J.P 2006).

2.2.1.2 Estimation des formants par prédiction linéaire

L'analyse par prédiction linéaire est l'une des méthodes les plus populaires pour extraire l'information spectrale du signal. La prédiction linéaire s'adapte au modèle tout pole pour les signaux de parole voisés. Les paramètres du modèle permettent de trouver les positions des formants, et il s'agit donc d'une technique paramétrique d'estimation des formants.

Les formants peuvent être estimés de deux façons différentes :

-Soit en suivant la méthode analytique, basé sur le calcul des pôles de la fonction du transfert du conduit vocal de (Markel.J D 1976).

-Soit graphiquement en détectant les pics qui peuvent apparaître sur l'enveloppe spectrale résultant de la prédiction linéaire du signal (Peak-picking) de (McCandless.S.S 1974).

- **Technique d'estimation des formants proposée par (Markel.J D 1976)**

Les formants correspondent mathématiquement, aux paires de pôles de la fonction de transfert du conduit vocal considérée comme un filtre linéaire. Chaque paire de pôles conjugués de la fonction de transfert notée H(z), représente un formant. A chaque pôle correspond une paire de fréquence-bande $\langle f_i, b_i \rangle$, f_i étant la fréquence de formant et b_i est la bande passante correspondante.

Un pôle présente un formant s'il vérifie des conditions. La première condition sert à vérifier la stabilité des pôles de la fonction de transfert dans le cercle unité. Plusieurs algorithmes ont adopté cette solution comme la méthode de (Snell.Roy C 1993) et la méthode de (Álvarez.A 1997).

La figure 2.2 ci-dessous décrit la démarche d'estimation des formants utilisant LPC :

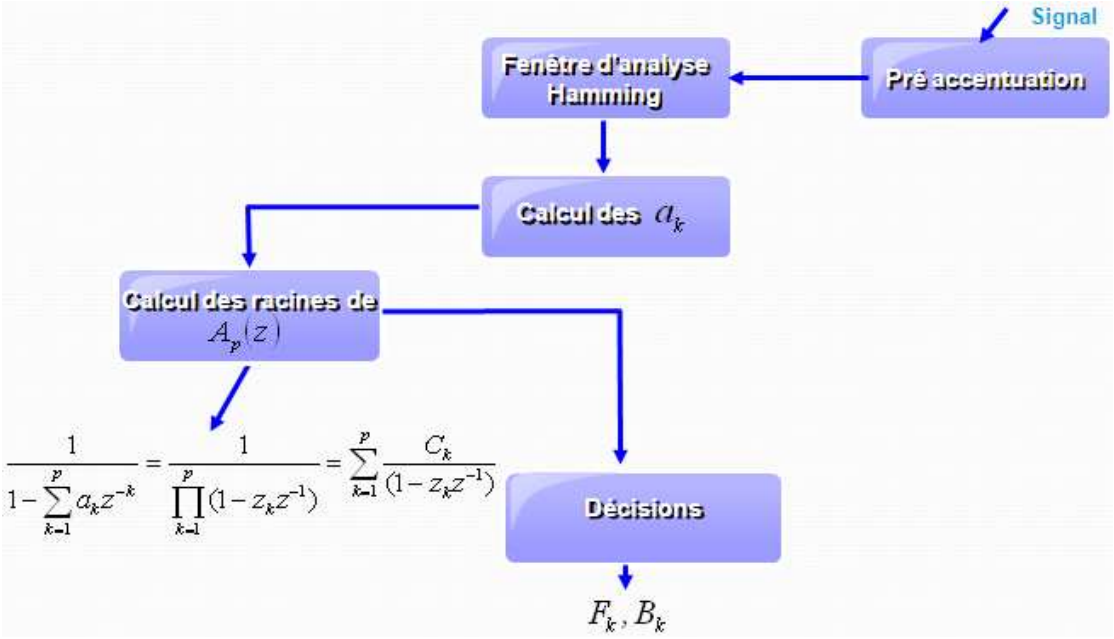


Figure 2. 2 : *Diagramme d'estimation des formants utilisant LPC*

Le signal est d'abord préaccentué pour égaliser les hautes fréquences qui sont moins énergétiques que les basses fréquences. Le découpage du signal en trames provoque des discontinuités aux frontières des trames qui se manifestent par des lobes secondaires dans le spectre. Pour compenser ces effets de bord, chaque trame est fenêtrée par une fenêtre de pondération. La fenêtre la plus utilisée est souvent la fenêtre Hamming avec une durée de 10 à 30 ms (Chaari.S 2005)**.** Le calcul des coefficients a_k est effectué en général par la méthode d'autocorrélation qui est plus populaire et plus efficace que la méthode de covariance pour l'estimation des coefficients du modèle LPC.

Les formants correspondent mathématiquement, aux paires de pôles calculés de $A_p(z)$ qui est considéré comme un filtre linéaire. Chaque paire de pôles conjugués représente un formant. On estime en général que la parole présente un formant par kHz de bande passante. A chaque pôle correspond une paire de fréquence-bande $\langle F_k, B_k \rangle$, F_k étant la fréquence de formant et B_k est la bande passante correspondante. Ainsi, le pôle calculé présente un formant s'il vérifie des conditions. La première condition sert à vérifier la stabilité des pôles de la fonction de transfert dans le cercle unité.

A une racine z_k proche du cercle unitaire correspond un certain formant F_k :

$$z_k = \mu_k e^{i\theta_k} \qquad \text{(Eq.2.8)}$$

L'estimation du formant F_k sera :

$$F_k = f_e \times \theta_k / 2\pi \qquad \text{(Eq.2.9)}$$

Avec f_e est la fréquence d'échantillonnage.

Cet algorithme est exécuté raisonnablement bien pour les sons purs et réguliers tels que les voyelles. Mais, le problème de fusion des fréquences estimées avec de fausses fréquences reste toujours sensible. Généralement ce sont les pics du bruit du fond, ce qui amène à une estimation incorrecte des formants. Etant donné que l'ordre de prédiction linéaire à utiliser conditionne le nombre de formants à prendre en compte, pour surmonter ce problème, il est préféré d'utiliser l'ordre de prédiction de 8 à 16 (Sundaram.N 2003). La

performance de cet algorithme est mauvaise pour les sons autres que les voyelles tels que les voyelles nasals et les segments non voisés.

- **Technique d'estimation des formants proposée par (McCandless.S.S 1974)**

Dans l'algorithme proposé par (McCandless.S.S 1974) qui est un algorithme de suivi de formants très utilisé, les trois premiers formants sont estimés à partir des pics du spectre de la prédiction linéaire du signal.

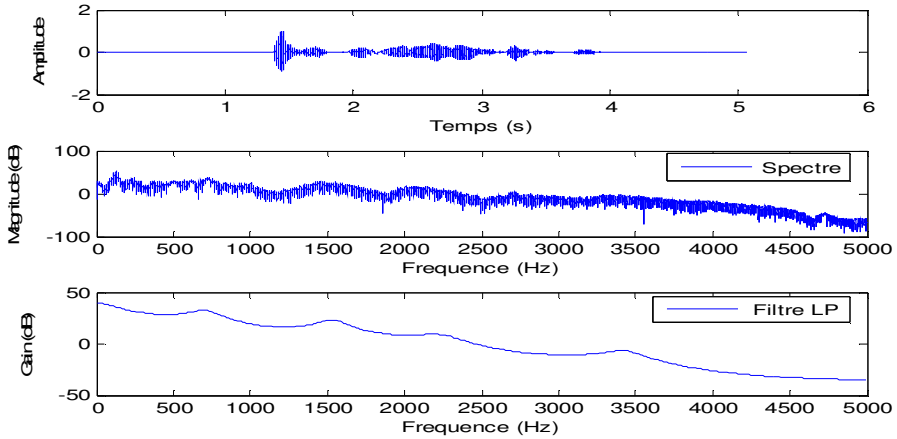


Figure 2. 3 : *Schéma du filtre de la prédiction linéaire et du spectre d'un signal de parole*

Il a été montré que la réponse fréquentielle du filtre est similaire à la version lissée du spectre d'amplitude de la transformée de Fourier du signal. L'estimation des fréquences des formants doit par conséquent être possible par une recherche des maxima du spectre du modèle (Voir Fig.2.3). Cette méthode d'estimation des formants est adoptée dans plusieurs algorithmes de suivi tel que la méthode de (Sundaram.N 2003)**.**

D'autres approches sont proposées par (Bruce. I C 2002) et (Chen.B 2004) et la méthode de (Mustafa.K 2006)**,** pour estimer les formants dans un environnement bruité et cela en appliquant un préfiltrage adaptatif avant de passer à l'estimation par l'analyse LPC.

2.2.2. Estimation des formants par lissage cepstral

Le lissage cepstral est une méthode non paramétrique qui vise à effacer les effets glottaux sur le spectre fréquentiel pour obtenir l'enveloppe spectrale correspondant à la réponse fréquentielle du conduit vocal. L'intérêt de l'analyse cepstrale est la déconvolution a posteriori du signal pour connaître la contribution de l'excitation glottique et les résonances du conduit vocal.

L'algorithme implémenté par (Schafer.R W 1970) est une technique non paramétrique d'estimation des fréquences des formants qui comporte le calcul du spectre lissé à travers le lissage cepstral et puis la détection des formants. Le cepstre est défini comme une transformée inverse du logarithme de la transformée de Fourier du signal; c'est la transformation homomorphique inverse (Schafer.R W 1970). La partie basse du cepstre peut être fenêtrée ou liftrée pour supprimer les pulsations glottales et extraire l'enveloppe spectrale du signal. Celle-ci dépend du conduit vocal et met en évidence les pics correspondants aux fréquences des formants. En effet, les fréquences des formants sont estimées à travers la sélection des pics de l'enveloppe spectrale obtenue à partir du lissage cepstral. Généralement, on estime seulement les trois premiers pics qui correspondent aux trois premiers formants.

Les techniques de lissage cepstral ne sont pas robustes dans le bruit blanc additif gaussien (AWGN) et donnent des résultats médiocres en présence d'autres locuteurs comme bruit de fond. L'algorithme ne donne pas non plus de bons résultats pour les locutrices même sans bruit.

L'analyse cepstrale est aussi utilisée pour la reconnaissance automatique, comme paramétrisation spectrale: coefficients MFCC (Mel Frequency Cepstrum Coefficients), coefficients LFCC (Linear Frequency Spectrum Coefficients), coefficients LPCC (Linear Predictive Cepstrum Coefficients) (Alessandro.C 1992) mais en intégrant un filtrage perceptif.

La méthode proposée par (Shahidur Rahman.M 2005) est aussi basée sur la transformation homomorphique; elle consiste à faire une étude détaillée d'estimation des formants de la parole de pitch élevé par la prédiction homomorphique. Cette méthode consiste à trouver les coefficients d'autocorrélation en appliquant l'analyse de la prédiction linéaire après avoir fait la transformation homomorphique du signal pour séparer entre les harmoniques de la source et la réponse impulsionnelle du conduit vocal. C'est ainsi que le résultat d'estimation de la réponse impulsionnelle du conduit vocal par l'analyse LP sera indépendant des variations de la fréquence fondamentale F0. En effet, dans la méthode d'autocorrélation conventionnelle de la prédiction linéaire (CALP) lorsque des segments fins sont extraits sur des différentes périodes de pitch, la séquence d'autocorrélation obtenue est en fait une version déformée d'autocorrélation de la réponse impulsionnelle du conduit vocal. Cela est dû aux répliques d'autocorrélation de la réponse impulsionnelle, répétée périodiquement avec une période équivalente à celle de pitch, qui déforment la fonction d'autocorrélation qui par conséquent empêchent une estimation raisonnable des formants. La

solution proposée par (Shahidur Rahman.M 2005) consiste à déconvluer la réponse impulsionnelle du conduit vocal du signal de parole en utilisant le filtrage homomorphique. La réponse impulsionnelle déconvoluée permet d'avoir une bonne estimation de l'autocorrélation et par concéquent une bonne estimation des formants.

L'utilisation de l'analyse cepstrale en combinaison avec la prédiction linéaire est appelée prédiction homorphique (Voir Fig2.4). Dans cette méthode, Shahidur vise à estimer les formants et appliquer la prédiction homorphique dans la perspective d'éliminer les effets dûs à une F0 du signal de parole élevée. On note qu'une solution raisonable des coefficients autoregressifs peut être obtenue uniquement si on obtient une bonne déconvolution de la réponse impultionnelle du conduit vocal. Pour cela, la méthode proposée recherche d'abord une estimation de la phase minimale du cepstre de la réponse impulsionnelle en utilisant le filtrage homorphique (Voir Fig2.4).

En calculant le spectre à différentes valeurs de F0, une déviation claire est observée pour le spectre CALP par rapport au spectre pur de la réponse impulsionnelle estimé par la méthode de (Shahidur Rahman.M 2005) lorsque F0 est égale à 250 Hz ce qui confirme l'effet de la distorsion de la fonction d'autocorrélation de CALP pour un pitch élevé.

L'ordre de LPC utilisé est 12 et la fenêtre d'analyse utilisée est Hamming d'une taille de 20 ms. Le signal est préccentué par un filtre du premier ordre. La taille de la fenêtre de liftrage utilisée est égale à 0.6P (P est a période de pitch) avec la valeur de F0 jusqu'à 250 Hz et la taille de la fenêtre vaut à 0.7P pour des valeurs de F0 plus garndes.

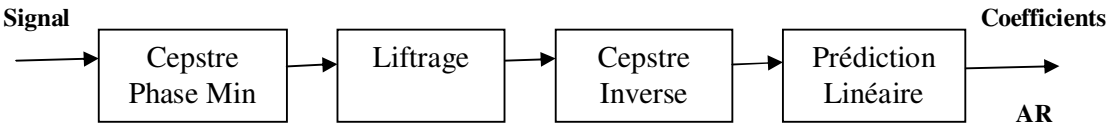


Figure 2. 4 : *Diagramme de la méthode prédiction homomorphique proposée par* (Shahidur Rahman.M 2005)

En testant les deux méthodes, celle proposée par (Shahidur Rahman.M 2005) et la méthode CALP sur la voyelle synthétique japonaise /e/ à différentes valeurs de F0, on remarque que les spectres estimés par la méthode proposée sont bien mis en évidence même à des valeurs de F0 élevées. Par contre les spectres estimés par CALP se déforment et se chevauchent à des F0 élevées.

L'estimation des fréquences des formants est obtenue en calculant les racines du dénomirateur de la fonction de trasfert définie par les coefficients de prédiction. En comparant la méthode proposée, prédiction homomorphique, avec la fonction d'autocorrélation CALP en les testant sur cinq voyelles Japonaises synthétiques avec différentes valeurs de F0, Shahidur trouve que le taux d'erreurs estimé pour les trois premiers formants est très inferieur à celui de CALP. En effet, le modèle conventionnel de l'analyse de prédiction linéaire seule est incapable de donner une estimation des formants précise à cause de la fonction d'autocorrélation utilisée pour estimer les coefficients autorégressifs qui peut être instable à cause de la périodicité du signal de parole.

En revanche, les coefficients autorégressifs trouvés par la prédiction homomorphique sont stables. La fiabilité de cette méthode est donc toujours supérieure à celle des autres méthodes courantes pour plusieurs applications en analyse de la parole.

2.2.3. Estimation des formants basée sur les modèles auditifs

L'efficacité du système auditif humain pour comprendre la parole en présence de bruit et d'autres conditions adverses a été une source d'inspiration dans plusieurs systèmes de traitement de parole. En effet, l'oreille est un organe qui est caractérisé par des échelles de traitement non linéaires et bien adaptée à la parole. Ainsi les basses fréquences sont perçues de manière plus fine par l'homme que les hautes fréquences. Plusieurs échelles essaient de rendre compte de cette non linéarité. Les échelles fréquentielles Mel et Bark (Calliope 1989) issues d'études psycho acoustiques qui déjà été intégrées avec succès dans les chaînes d'analyse. Ces échelles reproduisent approximativement la sensibilité de l'oreille.

Depuis quelques années, l'utilisation des modèles auditifs pour l'estimation des formants a suscité un nouvel intérêt d'autant plus que cela apporte une amélioration dans les applications de la reconnaissance de parole, spécialement en milieu bruité, par rapport aux autres méthodes classiques telles que LPC, lissage cepstral, MFCC, PLP et RASTRA-PLP.

Le modèle physiologique du système auditif est organisé à l'image des trois parties de l'oreille ; interne, moyenne et externe. Le limaçon et la membrane basilaire sont localisés dans la partie interne de l'oreille, ils peuvent être modélisés comme une réalisation mécanique d'un banc de filtres. En effet, ce sont les cellules ciliées IHC (Inner Hair Cells) distribuées tout au long de la membrane basilaire qui sentent des vibrations mécaniques et les

convertissent en impulsions électriques dans les fibres nerveuses qui à leur tour émettent des impulsions neurales dans le nerf auditif.

Inspiré de cette idée, un algorithme d'estimation des formants pour la parole bruitée proposé par (Metz.S W 1991) basé sur un modèle auditif connu utilise un ensemble d'histogrammes à intervalles EIH (Ensemble Interval Histogram). Dans la littérature, il y a aussi d'autres systèmes auditifs qui ont été conçus pour l'estimation des formants et qui vont être présentés dans cette section.

- **Système auditif proposé par (Metz.S W 1991)**

Le modèle auditif proposé par Metz est assimilé au modèle de l'oreille interne, il est construit à l'aide d'un banc de filtres passe bande BPF (comme ceux du limaçon). Les fréquences centrales des filtres successifs sont localisées uniformément tout au long de l'échelle Bark. L'implémentation du filtre consiste à des filtres IIR (Infinite Impulse Response) connectés en cascade et des filtres FIR (Finite Impulse Response). Les positions des pôles et des zéros sont réglées manuellement pour obtenir la réponse propre du modèle. Un total de 61 filtres est utilisé pour couvrir une largeur de bande de 200Hz à 3.2KHz. Les bancs de filtres BPF seront suivis par des détecteurs de passage de niveau LCD (Level Corssing Detector) avec un nombre de niveau égal à 8. Ces détecteurs modélisent la conversion du son en information neuronale. Ils sont distribués uniformément sur une plage dynamique de 40 dB sur une échelle logarithmique. Les sorties positives (sous forme d'intervalles) des détecteurs de passage à niveaux LCD sont inversées et placées dans un histogramme de fréquence (FH). La fenêtre d'analyse du signal est limitée à 20 intervalles. L'ensemble des FH (chaque FH correspond à un LCD) est sommé pour chaque BPF pour donner un EIH (Daoudi.K 2004) (Voir Fig.2.5).

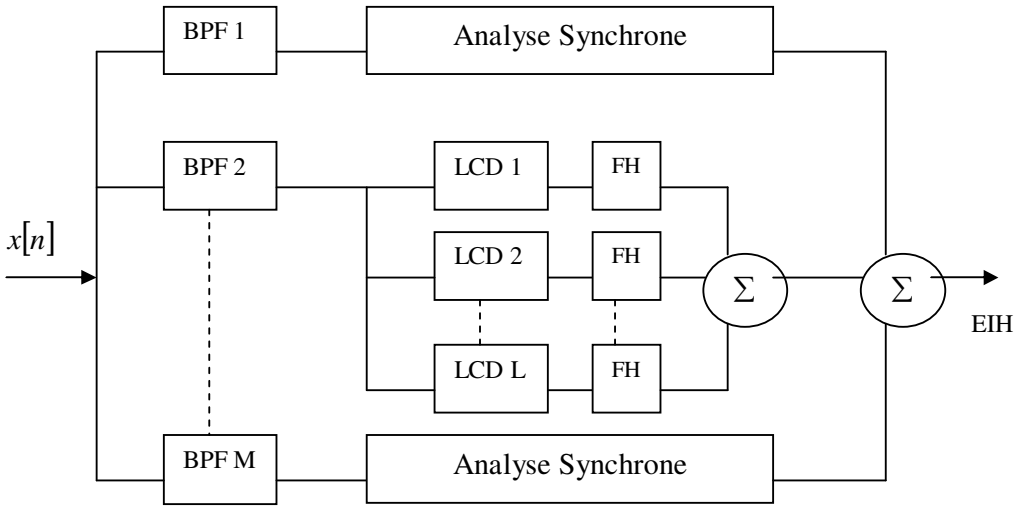


Figure 2. 5 : *Schéma du modèle auditif utilisé par* (Metz.S W 1991)

Cet algorithme a été implémenté en utilisant les pics du spectre lissé d'EIH pour l'estimation des fréquences des formants des sons voisés. Notons que les trois premiers formants du segment correspondent aux trois pics les plus élevés du spectre de l'EIH. Dans ce système d'estimation des formants proposé par Metz dans (Metz.S W 1991), le signal est segmenté en plusieurs segments successifs de 40ms avec un chevauchement de 75%. En testant cet algorithme sur une expression de parole synthétique, les auteurs ont remarqué que les pics des formants sont clairement mis en évidence sur le spectre d'EIH même avec un SNR de 0B.

Les auteurs ont testé l'algorithme sur quatre signaux de parole naturelle en ajoutant à un bruit blanc gaussien à de différentes valeurs de SNR (de 50dB à -10dB). Ils ont remarqué à travers les résultats qu'il y a d'autant plus de pics parasites que le SNR diminue. Le nombre de segments pour lesquels les formants ne sont pas détectés par l'algorithme augmente lorsque le SNR diminue.

- **Système auditif proposé par (Shamma.S 1988)**

Une autre extension qui a été introduite comme modèle du système auditif périphérique est l'inhibition latérale (LIN), proposée par (Shamma.S 1988). Elle correspond à une suppression de l'activité des fibres nerveuses sur la membrane basilaire causée par l'activité des fibres adjacentes. Ce phénomène est utilisé pour améliorer la robustesse au bruit en masquant la fréquence inhibé par le phénomène de LIN qui est donc considérée comme du

bruit. En effet, le modèle LIN est basée sur l'inhibition de chaque sortie des filtres par une ou plusieurs sorties des filtres voisins. Cela permet de rehausser les pics spectraux et d'améliorer la résolution fréquentielle. La sortie de chaque canal est calculée en soustrayant une moyenne pondérée de la somme de ses voisins suivie d'une opération de seuillage et d'une moyenne sur une fenêtre temporelle. De cette façon, les formants sont renforcés par la détection des filtres qui présentent une grande différence de phase avec leurs voisins. Il s'agit d'une approche simple, rapide et efficace pour la détection de la synchronisation et la production d'une représentation robuste des formants.

- **Système auditif proposé par (Seneff.S 1988)**

Une autre extension qui a été introduite dans les modèles du système auditif périphérique proposée par Seneff dans (Seneff.S 1988). Cette extension consiste à un détecteur de synchronie généralisée (GSD) comme modèle auditif au lieu de l'inhibition latérale utilisée par Shamma.

Le détecteur de synchronie généralisée implémente la propriété de verrouillage de phase « phase-locking » des fibres nerveuses dans le but d'augmenter la pertinence des pics spectraux dus aux résonances du conduit vocal. Il est conçu pour rehausser les formants, améliorer la résolution spectrale, réduire les caractéristiques du spectrogramme dû à l'excitation glottale et normaliser l'amplitude.

Pour chaque canal du banc de filtres, on a une sortie de synchronisation GSD. Ce dernier détecte la périodicité de la réponse temporelle en calculant l'autocorrélation de la sortie de chaque filtre en générant le rapport de la moyenne de la somme et de la différence des sorties de chaque filtre avec la version retardée de celle-ci. Le retard de chaque GSD doit correspondre à la fréquence centrale du filtre.

Le modèle GSD présente plusieurs avantages par rapport à d'autres systèmes auditifs. Tout d'abord il mesure la périodicité plutôt que la fréquence, la détection de la périodicité rend le système plus robuste pour un signal bruité. En outre, le calcul du rapport entre la somme et la différence des sorties des filtres effectue une normalisation de l'énergie qui réduit les fluctuations temporelles de la réponse dues à l'excitation glottale. Néanmoins, le modèle GSD présente aussi des inconvénients. En effet, la réponse de GSD contient d'importants pics parasites qui sont dus aux harmoniques de la fréquence fondamentale F0 en particulier pour les locutrices. Cela a été décrit par (Seneff.S 1988) comme un problème majeur qui limite l'efficacité du modèle GSD dans les applications en reconnaissance automatique de la parole.

En outre, il exige une adéquation précise entre les fréquences centrales des filtres et le temps de retard du GSD. Si cette précision n'est pas atteinte, des pics parasites apparaissent.

Pour surmonter les limites du modèle GSD citées ci-dessus, Abdelatty (Abdelatty Ali.A M 2002) (Abdelatty Ali.A M, 2000) a développé un nouveau système auditif appelé ALSD (Average Localized Synchrony Detection) à partir du modèle GSD.

- **Système auditif proposé par (Abdelatty Ali.A M,2002)**

Le système auditif ALSD proposé par Abdelatty dans (Abdelatty Ali.A M, 2002) est conçu pour corriger certaines limites du GSD. En fait, la sortie de chaque ALSD est la moyenne des GSD du canal i et de ses voisins. Le nombre de voisins considérés est décidé de manière empirique à partir de la résolution et de la bande passante des filtres utilisés. (Voir Fig 2.6).

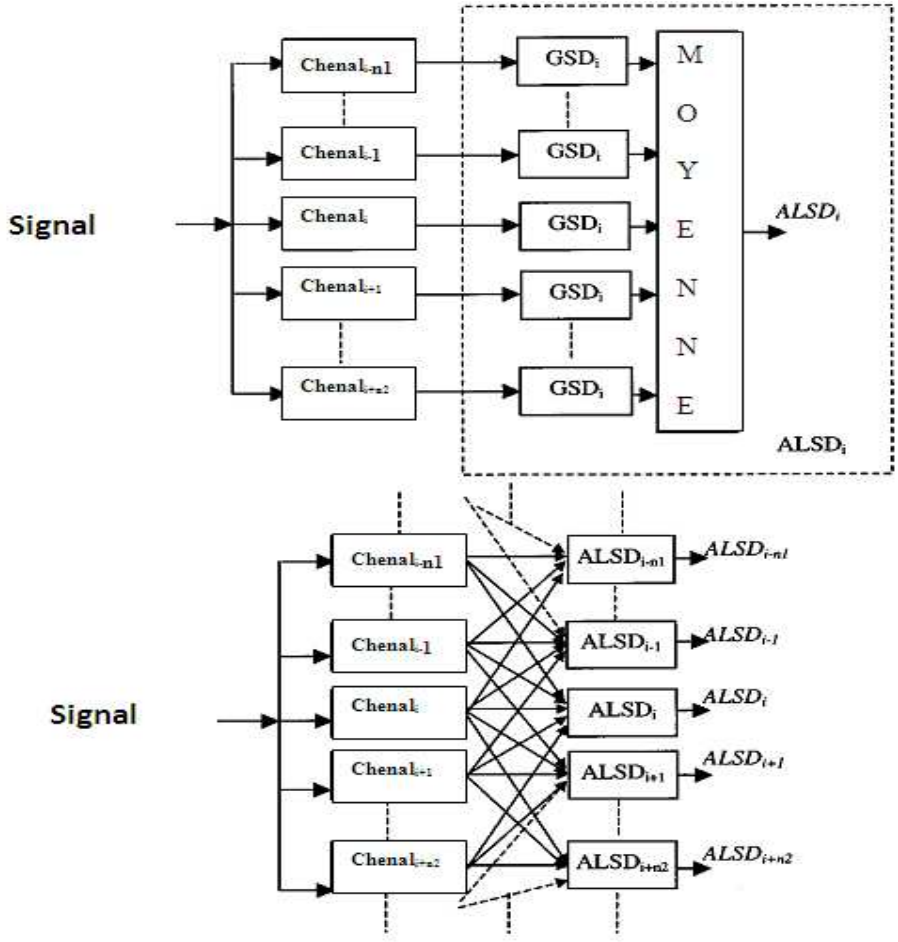


Figure 2. 6 : *Le diagramme du modèle ALSD proposé par* (Abdelatty Ali.A M 2002)

Le modèle auditif proposé par (Abdelatty Ali.A M 2002) détecte la périodicité dans le signal au niveau des filtres perceptifs tout en réduisant les pics parasites et la sensibilité à l'implémentation. Il incorpore aussi le filtrage à bande critique, la compression non linéaire de l'énergie, la modélisation de cellules ciliées et l'adaptation à court terme. La structure générale du système auditif proposé par (Abdelatty Ali.A M, 2002) est présentée par le shéma ci-dessous (Voir Fig 2.7).

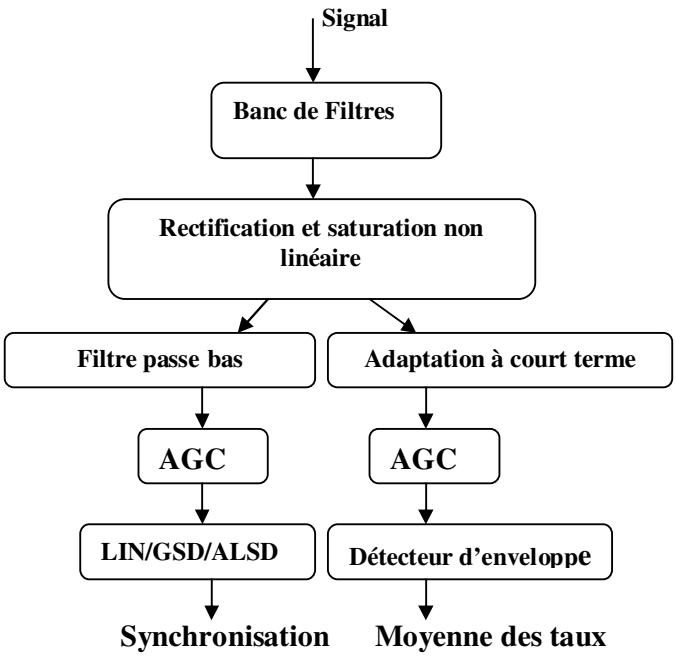


Figure 2. 7 : *Schéma du modèle auditif utilisé par* (Abdelatty Ali.A M 2002)

Le signal de parole est tout d'abord préfiltré à travers un banc de 36 filtres passe bande à l'échelle Barck. Le banc de filtres utilisé est un outil de simulation d'une cochlée analogique réelle qui a été mise en œuvre à l'aide de la technologie VLSI (Very Large Scale Integration). Le banc de filtres est suivi d'une étape non linéaire qui effectue une rectification demi-onde avec une compression et saturation non linéaire. Ensuite, le système est divisé en deux branches : la première à droite (voir Fig.2.7) qui donne la réponse moyenne des taux de sorties et la seconde à gauche (voir Fig.2.7) qui donne la réponse synchronisée.

La branche de la réponse moyenne des taux commence avec une adaptation à court terme et un module de masquage (STA) suivi d'un module de contrôle automatique des gains (AGC) et enfin se termine par un détecteur d'enveloppe (ED). Le détecteur d'enveloppe appliqué en sortie des niveaux précédents joue un rôle important pour capturer les changements rapides de la parole. Il permet de détecter les pics spectraux correspondants aux

formants en utilisant la moyenne des taux des sorties (Mean Rate). En effet, même au niveau du modèle auditif humain le principe du calcul de la moyenne des taux de sorties existe physiquement, ce phénomène consiste à capturer les informations essentielles extraites par le limaçon pour percevoir les pressions ondulatoires. Il couvre les toutes premières étapes du processus auditif qui visent à extraire des informations pertinentes pour la perception telles que les formants.

La particularité de ce nouveau système auditif est l'élimination significative des pics parasites tout en ajoutant une étape d'adaptation à court terme (STA). L'adaptation à court terme a été ajoutée pour améliorer la robustesse au bruit. Ce principe est d'ailleurs aussi utilisé dans le traitement RASTA et il a montré qu'il améliorait nettement la robustesse du système au bruit. De plus, l'adaptation et la prise en compte du masquage contribuent de manière significative au renforcement des frontières entre les différents segments de la parole. Le module de contrôle du gain (AGC) permet quand à lui de traiter des signaux d'amplitude variable au cours du temps. Enfin, le détecteur d'enveloppe (ED) est un simple filtre passe bas avec une fréquence de coupure égale à 50Hz.

D'autre part, la deuxième branche de l'algorithme qui calcule la réponse synchronisée est subit d'un filtre passe bas (LPF), un module de contrôle automatique de gain (AGC) et un détecteur de synchronie.

Le filtre passe bas qui est utilisé pour modéliser la suppression synchronisée qui se produit à des fréquences élevées en raison du temps de latence des neurones et de la réponse de nervosité. Suite à cette étape, il y a l'application du module de contrôle automatique de gain (AGC) et le détecteur de synchronie qui consiste à détecter les caractéristiques temporelles de verrouillage de phase de la réponse. Le détecteur de synchronie utilisé ici est ALSD qui peut être comparé aux deux autres détecteurs de synchronie classiques qui sont les modèles LIN et GSD. Ces modèles de sortie (ALSD, LIN et GSD) sont utilisés pour calculer les synchronies des sorties données par le contrôle automatique du gain AGC. Chaque sortie AGC est appliquée à un modèle de sortie réglé à la fréquence centrale correspondant au filtre auditif. Ainsi s'il y a un pic important dans le signal, il apparaîtra comme une périodicité au niveau de l'AGC.

Les résultats de détection des deux premiers formants ont été comparés pour ce qui concerne la méthode de la moyenne des taux de réponses avec le détecteur d'enveloppe et celle de la synchronie de réponses en utilisant les trois modèles auditifs ALSD, GSD et LIN.

Le système auditif ALSD a montré davantage de robustesse pour l'estimation des formants dans un environnement bruité que les deux autres modèles. Il s'est monté meilleur aussi par rapport à la méthode de la moyenne des taux de réponses.

- **Système auditif proposé par (Gläser.C 2010)**

Un autre type de modèle auditif important qui a été proposé ces dernières années par Patterson (Patterson.R.D,2004) pour l'estimation des paramètres du signal : c'est le banc de filtres gammatone. Les filtres gammatone sont souvent utilisés pour la modélisation du traitement du son dans la cochlée. Il effectue une analyse spectrale et convertit l'onde acoustique dans une représentation en multicanaux. Le filtre gammatone est défini dans le domaine temporel par sa réponse impulsionnelle :

$$gt(t) = at^{(n-1)} \exp(-2\pi bt)\cos(2\pi f_c t + \varphi) \qquad \text{(Eq.2.10)}$$

Où $(t > 0)$ et les premiers paramètres de filtre sont b et n : b est l'estimation de la durée de la réponse impulsionnelle, n est l'ordre de filtre, f_c est la fréquence de coupure et a est une constante de normalisation. Pour construire la totalité du banc de filtre. Pour cela, il faut utiliser le concept de largeur de bande rectangulaire équivalente (ERB) définie par :

$$ERB = 24.7(4.37 f_c / 1000 + 1) \qquad \text{(Eq.2.11)}$$

L'ensemble de ces deux équations citées précédemment définissent un banc de filtres auditifs gammatone. Lorsque le rapport f_c / b est élevé, comme il est dans le cas auditif, la bande passante du filtre est proportionnelle à b, et la proportionnalité constante ne dépend que de l'ordre du filtre, n. Par exemple, lorsque l'ordre est de 4, b est 1,019 ERB.

Une nouvelle technique qui repose sur le même principe est celle proposée par Gläser (Gläser.C 2010). Elle combine un prétraitement des principales fonctions du système auditif telle que l'utilisation d'un banc de filtres gammachirp pour l'estimation des formants et un algorithme probabiliste pour le suivi de formants. Cette combinaison est destinée à atteindre de meilleure performance pour estimer et de suivre les formants dans le bruit. L'architecture détaillée de cette méthode est présentée par la figure ci-dessous : (Voir Fig 2.8)

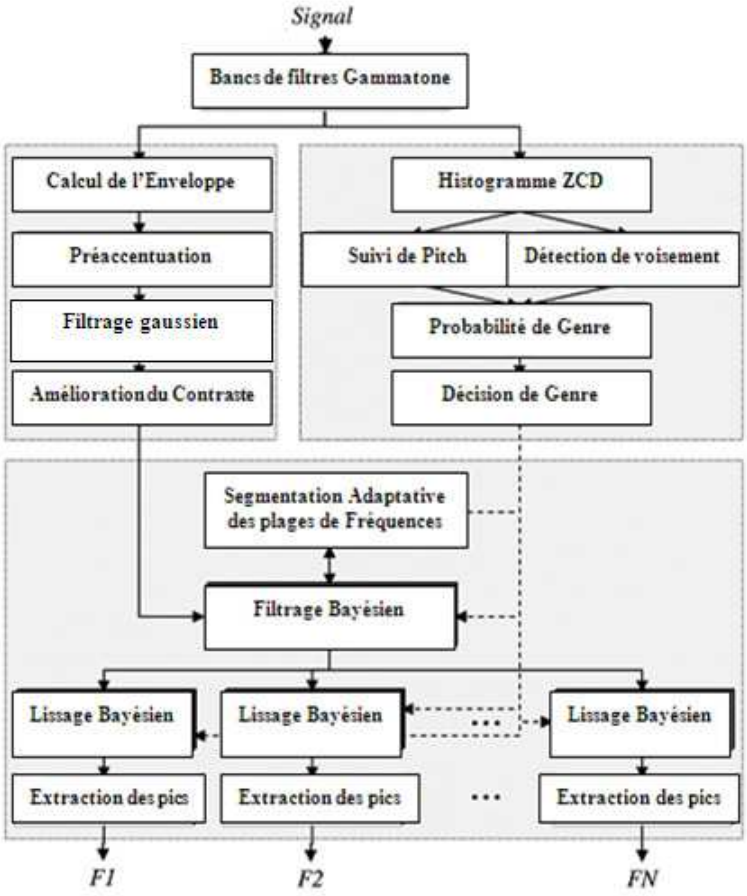


Figure 2. 8 : *Schéma de la méthode de suivi de formants proposée par* (Gläser.C 2010)

L'architecture de ce système de suivi de formants comporte trois blocs de traitement principaux. Le premier réalise un prétraitement auditif du signal d'entrée qui est ensuite filtré à l'aide d'une approche de filtrage gaussien et dont le contraste spectrographique des formants est renforcé. Le deuxième bloc est un algorithme de suivi de formants probabiliste basé le filtrage et le lissage bayésien. Le troisième est destiné à adopter le système au locuteur en détectant le sexe du locuteur. Il utilise un détecteur de la fréquence fondamentale et la détection du voisement.

Les trajectoires des formants dans le spectrogramme sont principalement affectées par deux causes. La première est la source d'excitation et le rayonnement qui introduisent une pente spectrale qui doit être corrigée par l'intermédiaire d'une préaccentuation. La deuxième est l'apparition de pics spectraux correspondants aux harmoniques plutôt qu'aux fréquences de résonance du conduit vocal dans le spectrogramme.

Pour palier ces difficultés, les auteurs utilisent un banc de filtres gammatone. C'est la première étape de l'algorithme qui a été schématisée dans le premier bloc de la figure ci-

dessus (voir Fig.2.8). Chacun des filtres couvre une plage de fréquence spécifique. 128 filtres sont utilisés pour analyser les fréquences de 80 Hz à 8KHz. Par la suite l'enveloppe spectrale est calculée par une rectification et un filtrage passe-bas. Les auteurs appliquent ensuite une préaccentuation de +6 dB/oct sur l'enveloppe spectrale en utilisant un filtre passe-haut de premier ordre qui élimine l'influence spectrale de l'excitation et du rayonnement. La structure des formants est aussi améliorée dans le spectrogramme à l'aide d'un lissage le long de l'axe de fréquences. L'application du filtrage des opérateurs gaussiens au spectrogramme préaccentué aplatit les harmoniques et renforce les valeurs spectrales des pics correspondant aux formants et réduit fortement l'énergie entre les formants.

Pour localiser les postions des formants et estimer leurs trajectoires, Gläser (Gläser.C 2010) a aussi effectué un filtrage bayésien associé à une segmentation adaptative des plages des fréquences. Cette étape de traitement a été schématisée dans le deuxième bloc de la figure ci-dessous (voir Fig.2.8). L'avantage du filtrage bayésien dans le traitement des signaux bruités a été montré dans plusieurs applications. Le filtrage bayésien appliqué dans cette méthode est adapté au suivi multi cibles (C'est la présence de plusieurs fréquences candidates à chaque instant pour suivre les trois premiers formants en même temps) ce n'est pas comme le filtrage bayésien standard.

Pour extraire les positions des formants, le filtrage bayésien est accouplé par l'algorithme de segmentation adaptative qui joue un rôle important dans la division des ensembles des fréquences candidates en parties spécifiques pour chaque formant. Même durant les fréquences des formants très adjacentes et proches, cette approche donne de bons résultats. Suite à cette étape pour assurer la continuité des trajectoires et obtenir des trajectoires de formants plus robustes d'autres traitements sont nécessaires. Au niveau du filtrage bayésien, les futures observations doivent être prises en compte, pour cela les auteurs ont appliqué un lissage bayésien. Le résultat du lissage bayésien appliqué montre que les trajectoires des formants sont nettement améliorées et cela est dû aux contraintes de continuité intégrant les anciennes et les futures observations et par conséquent la plupart des ambiguïtés sont résolues. Le calcul final des positions exactes des formants peut être fait en faisant l'extraction des pics des composantes lissées.

L'étape de l'extraction des formants et l'estimation des trajectoires peut être améliorée en prenant en compte d'autre information supplémentaire telle que la connaissance du sexe du locuteur en faisant l'extraction de pitch (voir le bloc3 à droite sue la Fig.2.8). Cette information peut être introduite dans la distribution des probabilités des formants ce qui est raisonnable parce qu'on sait bien que les profils des formants des hommes, des femmes et des

enfants diffèrent d'une façon significative. Suite à cela, la robustesse des trajectoires des formants est améliorée même à la variabilité des locuteurs. Cette méthode a été comparée avec d'autres méthodes existantes qui sont : les logiciels publics tels que Praat (Praat) et Wavesurfer (Wavesurfer) et la méthode de Mustafa (Mustafa.K 2006). Les résultats de la comparaison ont montré que la méthode proposée par (Gläser.C 2010) a présenté une robustesse significative aux différents types de bruits testés par rapport aux autres approches de suivi de formants.

Plusieurs méthodes d'estimation des caractéristiques spectrales basées sur les modèles auditifs ont été proposées dans la littérature pour obtenir une meilleure robustesse au bruit, mais il apparaît que le succès dépend de l'exactitude de la robustesse du modèle auditif utilisé.

2.3. Les techniques de suivi des trajectoires des formants

L'estimation des fréquences des formants à un instant ne suffit pas. Il faut en effet connaître l'évolution de ces fréquences dans le temps. Des efforts considérables ont donc été consacrés au développement d'algorithmes sophistiqués de sélection des fréquences candidates afin d'obtenir les trajectoires des formants. Dans la littérature, des méthodes utilisent la programmation dynamique, les modèles de Markov cachés (HMM) ou bien les modèles de mélange de gaussiennes (GMM) et depuis peu le filtrage de Kalman. Certaines parmi ces méthodes combinent plusieurs techniques de suivi. Nous allons présenter quelques unes de ces approches.

2.3.1. Techniques basées sur la programmation dynamique

L'approche par Programmation Dynamique fournit d'excellentes performances lorsqu'elle a été appliquée à l'estimation de la fréquence fondamentale (Secrest.B 1983) et on l'a donc testée aussi pour l'estimation des fréquences formantiques. Dans la littérature, plusieurs méthodes de suivi de formants appliquent la technique de la programmation dynamique comme celle de (Talkin.D 1987) qui utilise des contraintes de continuité entre fréquences formantiques. Une autre technique proposée par Xia (Xia.K 2000) plus récemment est basée sur le même principe que nous présenterons plus bas.

- **Approche proposée par (Xia.K 2000)**

La technique de (Xia.K 2000) opère deux étapes principales :

-La première étape est similaire à celle développée par (Talkin.D 1987) ; elle consiste à trouver l'estimation optimale des trajectoires de formants en imposant les contraintes de continuité en utilisant la programmation dynamique.

-La deuxième étape consiste à appliquer une étape de post traitement destinée à rendre l'estimation des formants plus robuste et à étendre l'estimation dans les régions nasales et obstructives. Pour cela, la programmation dynamique a été couplée avec l'information de segmentation du signal en trois régions principales qui sont : voyelles, nasales et obstructives.

Similaire à la technique de Talkin, l'algorithme de la programmation dynamique utilise les racines complexes de dénominateur de la fonction de transfert du prédicteur linéaire comme un ensemble de formants candidats. En choisissant un ordre de prédiction approprié, il est possible de réduire les problèmes de fusion de formants sans ajouter trop de candidats inutiles.

Dans cette méthode, le signal de parole est échantillonné à 10 kHz pour ne conserver que les trois ou quatre premiers formants inférieurs à 5KHz et préaccentué par un filtre passe haut du premier ordre. Un autre filtre FIR symétrique non causal est appliqué pour atténuer les composantes de très basses fréquences. Une analyse par la méthode LPC par autocorrélation est effectuée à chaque trame utilisant une fenêtre d'analyse cosinus d'une durée de 49ms. Les pôles sont obtenus en cherchant les racines du polynôme de prédiction linéaire d'ordre 12. Les pôles réels qui ont simplement contribué à la pente du spectre global sont éliminés. Les pôles complexes restants sont ensuite représentés par leur fréquence et largeur de bande $\{F_i, B_i\}$ avec i=1,2… N (N<=p/2) et p l'ordre de la LPC. Pour trouver le meilleur ensemble de trajectoires pour N formants à travers un treillis des formants candidats, on minimise le coût de connexion des fréquences des formants à chaque trame tout au long de l'analyse des trames en utilisant l'algorithme de Viterbi.

La fonction du coût de la programmation dynamique est défini par :

$$C(t,n) = C_{local}(t,n) + \min_m \{C_{tran}((t,n),(t-1,m)) + C(t-1,m)\} \qquad \text{(Eq.2.12)}$$

où $C(t,n)$ est le coût cumulatif au nœud (t, n). n représente un appariement des fréquences candidates aux formants à l'instant t et m un appariement à l'instant t-1. Chaque nœud de l'algorithme de Viterbi représente un appariement des pôles aux formants en d'autres termes un étiquetage des pôles en termes de formants.

$C_{local}(t,n)$ est le coût local à (t, n) qui reflète des connaissances concernant les formants sans le contexte temporel.

Trois types de traitements sont utilisés pour l'algorithme :

- Affiner les estimations initiales de formants pour essayer de faire en sorte qu'il y ait une information disponible pour tous les formants quand il n'y a pas assez de fréquences candidates manquantes. Les limites inférieures et supérieures des quatre premiers formants sont :

  100<F1<1500, 500<F2<3500, 1000<F3<4500 et 2000<F4<5000
- Réduire le coût de transition dans l'algorithme de programmation dynamique lorsque les fréquences des formants changent rapidement au cours de temps au moment des transitions.
- Étant donné que les fréquences des formants varient considérablement autour des valeurs neutres du conduit vocal (utilisées comme des estimations initiales) un coût linéaire est utilisé pour pénaliser la déviation des fréquences des formants par rapport aux valeurs neutres. Les valeurs neutres du conduit vocal sont données ainsi en Hz:

  Fn_1 =500, Fn_2 =1500, Fn_3 =2500 et Fn_4=3500

Pour déterminer les N premiers formants $C_{local}(t,n)$ est défini par :

$$C_{local}(t,n) = \sum_i \left\{ \alpha_i B_i^2 + \beta_i \left| F_i - Fn_i \right| / Fn_i \right\} \qquad \text{avec} \quad i = 1,2...N \qquad \text{(Eq.2.13)}$$

où $\{F_i, B_i\}$ est la paire fréquence-largeur de bande pour la i^{ème} composante du nœud (t,n).

$C_{tran}((t,n),(t-1,m))$ est le coût de transition du nœud $(t-1,m)$ au (t,n).

$$C_{tran}\left((t,n),(t-1,m)\right) = \sum_i \gamma_i \left(F_{in}(t) - F_{im}(t-1) \right)^2 \qquad \text{avec} \quad i = 1,2...N \qquad \text{(Eq.2.14)}$$

Puisque les formants varient lentement à l'intérieur des segments phonétiques, une fonction de coût quadratique de changement de fréquence inter-trame est utilisée pour

pénaliser les discontinuités entre $F_i(t-1)$ et $F_i(t)$ qui sont les fréquences du i^{ème} formant aux nœuds (t-1, m) et (t, n) respectivement.

Les constantes α_i , β_i et γ_i contrôlent la pondération relative des différentes fonctions de coût pour chaque formant. Ils sont déterminés empiriquement par des expériences supervisées sur 20 phrases de la base TIMIT.

Les auteurs ont testé cet algorithme avec quelques phrases de la base TIMIT et ont trouvé qu'il présente une bonne performance pour les segments de parole bien représentés par le modèle tout pôle LPC et spécialement pour les zones voisées du signal. Par contre, pour les sons nasalisés ou non voisés, les résultats se dégradent parce que les fréquences attendues ne sont pas trouvées comme formant potentiel. Plusieurs techniques de post traitement ont été proposées pour éliminer les estimations erronées et ne conserver que les bonnes estimations.

Les auteurs ont proposé une nouvelle stratégie utilisant la recherche de Viterbi en ajoutant l'information de segmentation du signal. Les formants sont fondamentalement différents dans les segments des voyelles, nasales et obstructives. Pour les voyelles, les formants sont souvent très marqués et la représentation avec le modèle tout pôle est raisonnable. Pour les nasales, les formants peuvent être faibles à cause de zéros très proches. Pour les obstructives, les formants sont souvent peu marqués ou absents. La donnée des formants dans les sons voisés est très utile à l'estimation des formants, plus pertinente et plus fiable qu'ailleurs. Par conséquent au lieu d'appliquer la recherche de Viterbi sur tout le signal, les auteurs segmentent d'abord le signal en trois régions : voyelles, nasales et obstructives soit à l'aide d'une segmentation manuelle ou bien à l'aide d'un système de reconnaissance en grandes classes phonétiques.

La recherche de Viterbi a été réalisée séparément pour les voyelles et les nasales uniquement là où on attend des trajectoires de formants continues.

Les trajectoires des formants peuvent présenter des interruptions dues à l'absence des pôles candidats ou des points aberrants. Pour résoudre cette ambiguïté, un lissage médian est d'abord effectué sur les premiers formants estimés pour écarter les valeurs aberrantes. Ensuite, chaque trajectoire de formants est en outre décomposée en concaténation de fragments lisses à partir d'un critère de continuité. Chaque saut de fréquence inter trame est comparée à un seuil à fin de déterminer les points de discontinuité. Le segment continu le plus long trouvé de cette manière est conservé comme formant d'ancrage. Ensuite, un faisceau de recherche locale est réalisé dans les deux directions commençant par le formant d'ancrage dans le but de trouver les formants absents dans cette région et dans les régions obstructives

voisines. Les fréquences des pôles constituant le formant d'ancrage sont prises comme valeurs de référence pour les segments voisins afin de rechercher des candidats donnant lieu à un saut fréquentiel minimal. Si aucun pole candidat n'a été trouvé, le faisceau de recherche s'arrête et la trajectoire des formants trouvée était la meilleure que l'on pouvait obtenir. Cette recherche à l'aide de la programmation dynamique minimise le coût cumulatif de l'ensemble de fréquences candidates ce qui donne un lissage des trajectoires des formants.

Pour évaluer les performances de cette technique, le suivi a été testé sur 34 phrases de différents locuteurs masculins choisis aléatoirement dans la base TIMIT. Les données de référence des formants ont été construites en utilisant un système de suivi de formants commercialisé « Entropic » (Xwaves 5.3.1) et en corrigeant les erreurs de suivi manuellement. Les transcriptions manuelles sont utilisées pour segmenter les phrases en trois catégories phonétiques.

En comparant l'algorithme proposé par Xia avec le suivi de formants de référence, les auteurs localisent les erreurs de l'algorithme proposé lorsque la différence entre la valeur manuelle des fréquences des formants et le résultat de l'algorithme dépasse 200 Hz. Le taux d'erreur est calculé séparément pour les voyelles et les nasales. Les régions obstructives n'ont pas été prises en compte puisque les données de référence ne sont pas disponibles. Les résultats des tests ont montré que le taux d'erreur est faible pour les voyelles et plus élevé pour les nasales. La plupart des erreurs correspond à l'inadaptation du modèle tout pôle. Une solution pour résoudre ce problème en partie consiste à augmenter l'ordre de la LPC quand les sauts en fréquence sont importants ce qui traduit le fait que certains pôles ont sans doute été oubliés. Il apparaît aussi que dans le cas ou le spectre change rapidement, la discontinuité des formants doit être pénalisée plus faiblement que pour les régions spectrales stables pour lesquelles la continuité des formants est en grande partie respectée.

- **Approche proposée par (Manocha.S 2005)**

La méthode de Xia et al (Xia.K 2000) a été reprise par Manocha. S (Manocha.S 2005) et améliorée sur plusieurs étapes principales de l'algorithme (voir Fig.2.9).

Puisque le suivi de formants est diffcile dans certaines conditions, les auteurs ont ajouté une mesure de confiance pour chaque formant à chaque instant. Ils ont aussi raffiné les estimations de formants initiales pour essayer de faire en sorte qu'une information plus complète soit disponible en combinant les analyses LPC d'ordre 12 et 16. L'algorithme favorise aussi les transitions formantiques fortes en réduisant le coût de transitions dans l'algorithme de programmation dynamique lorsque les formants changant rapidement. Etant

donné que les les formants varient lentement au cours de temps Manocha et ses collègues ont employé des fréquences de référence autoadaptatives plutôt que d'utiliser les valeurs neutres du conduit vocal utilisées pour l'estimation initiale. Les valeurs neutres sont donc remplacées par la moyenne des valeurs des formants observées.

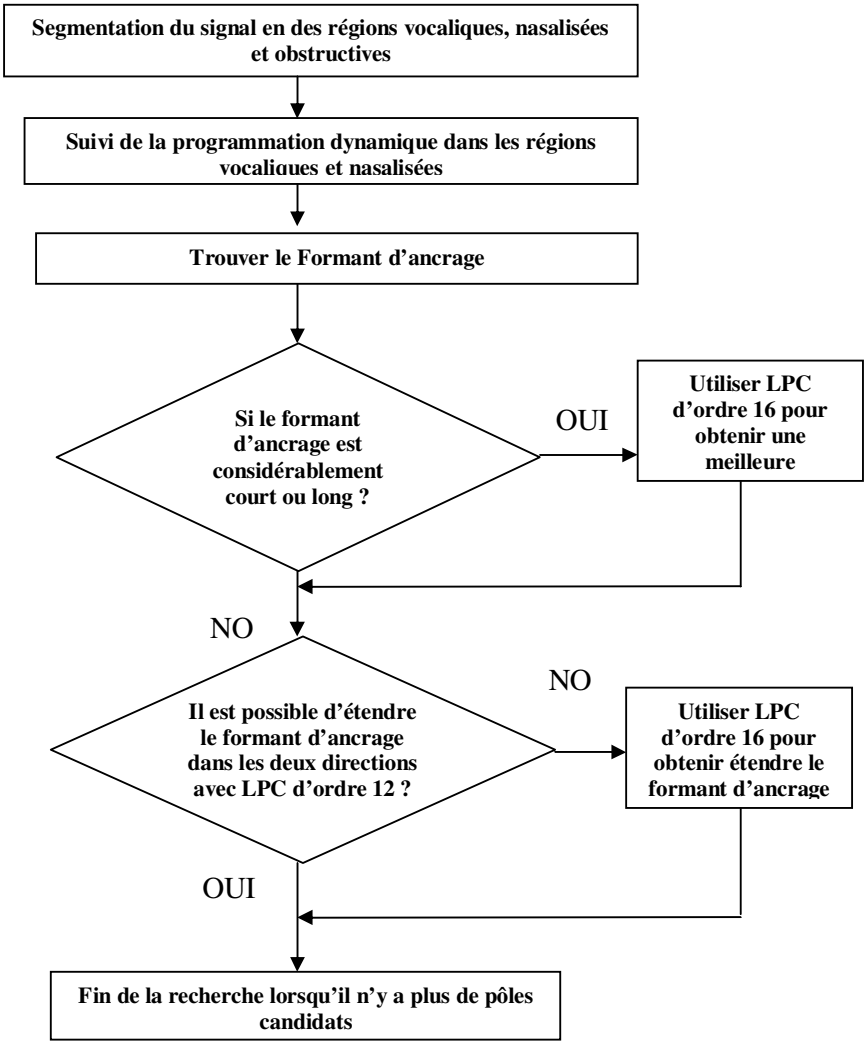


Figure 2. 9 : *Schéma de l'algorithme de suivi de formants proposé par* (Manocha.S 2005)

La mesure de confiance attribuée à chaque fréquence formantique augmente en fonction de :

-la continuité du suivi de formants

- l'amplitude élevée et la largeur de bande très basse d'un formant.

Les mesures qui réduisent le degré de confiance sont :

- L'existance de deux pics concurrents pour un seul formant

- L'écart de la moyenne par rapport à deux écarts type

Les suivis de référence sont calculés dans les régions voisines et sur la phrase complète. Les valeurs de confiance sont comprises entre 0 et 1. L'utilisation d'un seuil de confiance noté η permet de séparer les valeurs de confiance basses et élevées.

L'algorithme a été testé sur 72 phrases (33 locuteurs et 39 locutrices) prononcées par différents locuteurs sélectionnés aléatoirement dans la base de données TIMIT. Les valeurs de référence sont obtenues en utilisant l'algorithme public « Entropic » (Xwaves 5.3.1) de suivi de formants et d'apporter des corrections à la main si c'est nécessaire. Des transcriptions à la main de TIMIT sont faites par la segmentation des limites des différentes classes phonétiques. L'évaluation de l'algorithme a été faite seulement sur les régions des voyelles et avec différents seuils de confiance. Les résultats montrent de bonnes performnaces pour les locuteurs ainsi que les locutrices seulement pour les voyelles.

- **Approche proposée par (Lee.M 2005)**

La méthode de Lee, montre que si l'on connaît les sons du signal, le taux d'erreur de suivi de formants peut être significativement réduit (voir la Fig.2.11).

-La première étape consiste à segmenter le signal à l'aide d'un alignement temporel forcé (Sjölander.K 2003) (voir Fig.2.10) ; La transcription utilisée par cet alignement est obtenue à l'aide des niveaux linguistiques d'un système de synthèse de la parole.

- la signal est analysé à partir d'une analyse LPC d'ordre 12 sur tous les sons en utilisant une fenêtre de 25 ms. On affecte à chaque formant la valeur nominale en fonction de l'étiquetage phonétique.

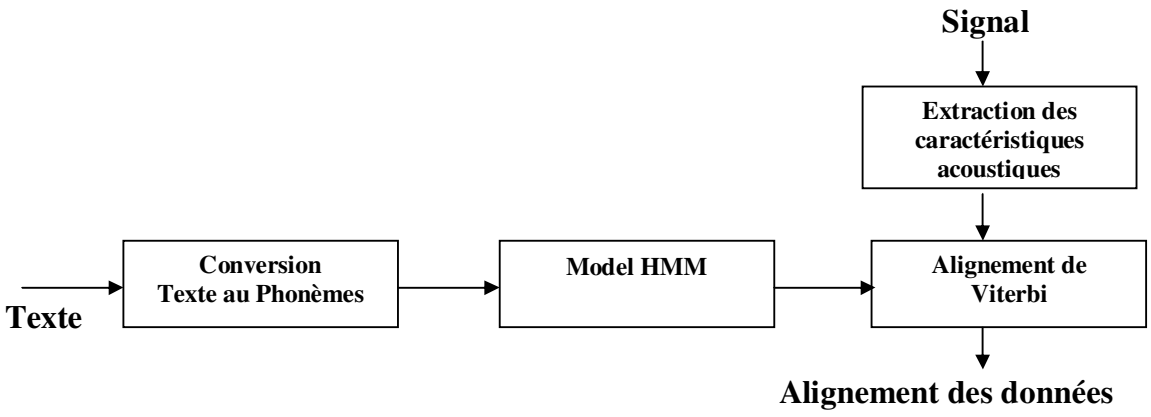


Figure 2. 10 : *Schéma de système de segmentation automatique basé sur les HMM* (Lee.M 2005)

Les suivis de formants nominaux pour tout le signal sont ensuite obtenus par interpolation des fréquences nominales. Dans le but de prendre en compte la coarticulation, différentes méthodes d'interpolation sont utilisées en fonction du contexte phonémique. Le processus d'interpolation garantit la robustesse de l'algorithme de suivi de formants vis-à-vis des erreurs provoquées par l'algorithme de segmentation. Dans cette méthode Lee a utilisé deux types d'interpolation qui dépendent des types de phonèmes traités. En général, on utilise l'interpolation linéaire sauf que pour certains phonèmes tels que liquides et sons nasalisés pour lesquels il a utilisé une interpolation non linéaire pour mieux approcher l'effet de coarticulation.

Finalement, un ensemble des formants est choisi à partir des formants candidats (les valeurs calculées par la LPC) de telle manière que le suivi de formants final couvre les suivis nominaux en satisfaisant les contraintes de continuité. Le choix des contraintes de continuité utilisées est fait selon les transitions phonémiques et l'énergie RMS du signal. En effet, si l'énergie du signal change brusquement lors des transitions rapides entre différents phonèmes on peut réduire le poids des contraintes de continuité parce que ces contraintes ne sont plus pertinentes.

Le suivi de formants obtenu est donc guidé par la connaissance des trajectoires nominales formantiques construites à partir de la connaissance de la suite des sons prononcés, ce qui rend l'algorithme plus performant.

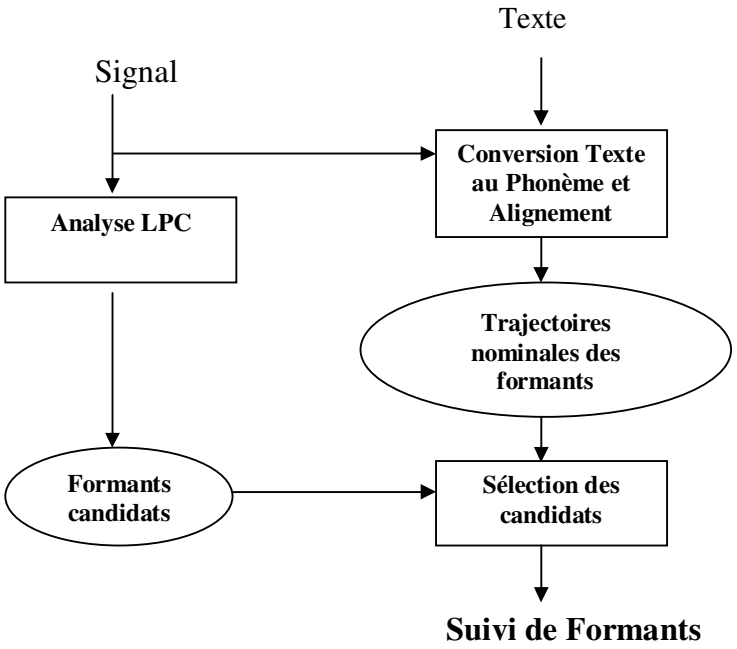


Figure 2. 11 : *Schéma global de l'algorithme de suivi de formants proposé par* (Lee.M 2005)

L'algorithme a été testé sur des expressions de parole naturelles et les résultats ont été comparés avec d'autres méthodes de suivi basées sur la programmation dynamique et utilisant seulement l'intégration des contraintes de continuité. Les résultats obtenus par la méthode de Lee, montrent une bonne performance malgré les erreurs de segmentation. En général, les erreurs de segmentation avec une trame de taille 40 ms ou moins ne provoquent pas souvent des dégradations critiques de la performance de l'algorithme car les valeurs nominales des formants proches des limites des sons sont interpolées correctement (sauf pour les nasales). Pour vérifier cela Lee a testé l'algorithme avec une segmentation manuelle et les résultats ont montré que cet algorithme restait toujours performant. Cet algorithme a réduit le taux d'erreur de suivi de formants pour les sons voisés à 5.03% pour les locuteurs et à 3.73% pour les locutrices par rapport aux autres méthodes traditionnelles (13% pour les locuteurs et 15.82% pour les locutrices).

2.3.2. Approches basées sur les modèles de Markov cachés

Il y a plusieurs années, Kopec a réussi à mettre en œuvre un algorithme de suivi de formants basé sur les HMM (Kopec.G 1986). Il a introduit deux nouveaux aspects pour améliorer l'algorithme de suivi de formants. Tout d'abord, il a utilisé deux modes de suivi, l'un pour le suivi d'un seul formant et le second pour plusieurs formants. Deuxièmement, l'algorithme forward-backward a été utilisé à la place de l'algorithme de Viterbi pour vérifier et évaluer les formants, car l'algorithme de Viterbi ne peut générer qu'une seule séquence d'état mais pas une distribution de probabilité (Gang.L 2010).

Il existe d'autres approches de suivi de formants basées sur les modèles de Markov telle que celle de (Acero.A 1999). Elle consiste à réaliser le suivi de formants et la synthèse de la parole en utilisant les HMM. Les vecteurs utilisés sont composés des trois premières fréquences de résonance détectées par LPC, leur bande passante et de leur évolution temporelle. L'algorithme de Viterbi est utilisé pour trouver le chemin le plus probable et imposer des contraintes de continuité en utilisant la distribution a priori de chaque son pour la sélection des formants candidats. Le suivi est lissé en appliquant un ajustement au formant brut lorsque l'erreur de l'estimation est grande relativement à la variance de l'état de HMM correspondant. Dans cette approche à base de HMM, Acero a adopté une synthèse pilotée par

les données presque complètement basé sur la modélisation des positions et des bandes passantes des formants.

Les méthodes les plus performantes actuellement sont celles qui intègrent le contexte phonémique tel que la méthode d'analyse des formants basée sur l'initialisation, l'apprentissage et la dépendance du contexte phonémique des HMM proposée par (Toledano.T.D 2006) que nous décrivons plus bas.

- **Approche proposée par (Toledano.T.D 2006)**

La méthode de (Toledano.T.D 2006) est la plus récente et elle repose sur celle de Lee. Elle utilise la segmentation phonétique automatique (Toledano.T.D 2003) au lieu de la segmentation basée sur les états. Ensuite, un algorithme de détection des formants candidats par l'analyse LPC est utilisé pour générer au plus six formants candidats par trame et enfin, un algorithme de Viterbi modifié est appliqué pour trouver les trajectoires des formants les plus probables. Les formants candidats et les contraintes de continuité sont imposés par la segmentation phonétique automatique. Le schéma présenté ci-dessous est un diagramme détaillé des différentes étapes de la méthode proposée. (Voir Fig.2.12)

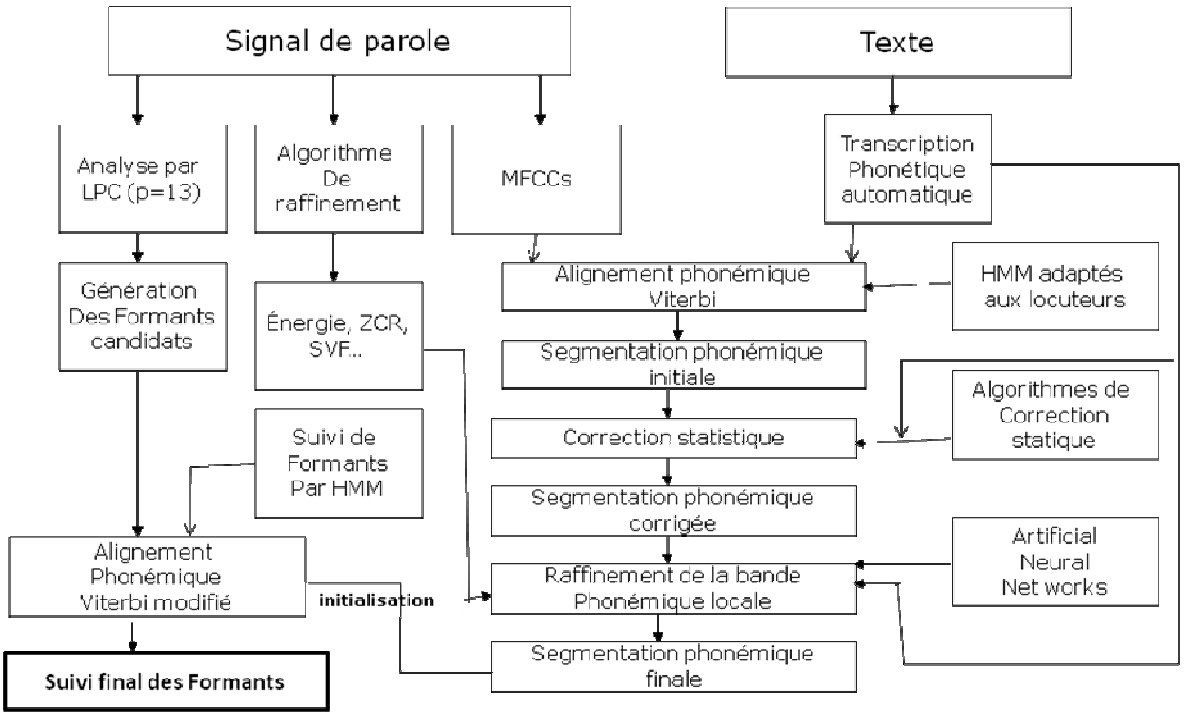


Figure 2. 12 : *Schéma détaillé de système de suivi de formants proposée par* (Toledano.T.D 2006)

L'algorithme automatique de segmentation phonémique présenté dans (Toledano.T.D 2003) est basé sur une approche de deux étapes. Ces deux étapes correspondent bien au premier bloc à droite décrit dans le schéma présenté ci-dessus. (Voir Fig.2.12)

- La première est basée sur un contexte phonémique modifié dépendant des modèles HMM adaptés aux locuteurs. Ces modèles sont basés sur un système de reconnaissance phonémique de la parole.

- La deuxième étape consiste à appliquer un algorithme d'amélioration de la segmentation basé sur les ANN (Artificiel Neural Networks) ainsi que d'autres dispositifs tels que le calcul de ZCR (Zero Crossing) et de la fonction de variation spectrale dont le but est de détecter les changements spectraux dans le signal de parole.

Les ANN réussissent bien à combiner l'information de segmentation fournie par les HMM et ces dispositifs. Ils fournissent donc une segmentation plus précise ce qui rend l'algorithme d'autant plus performant.

Le deuxième bloc à gauche consiste à générer des formants candidats en utilisant l'analyse LPC (autocorrélation) d'ordre 13. Tout d'abord, le signal de parole est ré-échantillonné à 8 KHz et découpé par la fenêtre de Hamming d'une longueur de 25 ms. L'étape de génération des formants candidats produit pour chaque trame trois vecteurs de formants F1, F2 et F3 contenants les fréquences candidates ainsi que trois autres vecteurs B1, B2 et B3 contenants leur largeur de bande. Le vecteur d'observation des HMM est composé alors de douze éléments : les trois vecteurs des formants, les trois vecteurs de leur largeur de bande ainsi que leurs variations en prenant en compte les vecteurs correspondant à la trame précédente. A partir de la séquence de vecteurs d'observation choisie, les HMM peuvent être entraînés en utilisant la méthode de Baum Welch. En itérant cette procédure plusieurs fois, le système proposé converge vers de bons modèles capables de sélectionner les bons formants candidats pour chaque trame et chaque phonème. Ensuite, un algorithme de Viterbi modifié est appliqué dans ce système pour permettre de faire l'alignement phonémique avec la séquence de formants candidats générée par la LPC et le suivi de formant fait par les HMM à chaque trame. La différence essentielle avec l'algorithme de Viterbi traditionnel est qu'à chaque état, on peut avoir différents vecteurs de candidats observés au lieu d'avoir un seul vecteur. La recherche de l'algorithme de Viterbi est limitée en utilisant la segmentation phonémique et en imposant des contraintes de continuité qui consistent à intégrer dans les

modèles HMM, l'information sur le phonème qui précède la transition courante et celle du phonème qui vient après.

Les auteurs ont étudié les performances du suivi en utilisant trois techniques différentes d'initialisation des modèles HMM (Initialisation uniforme, moyenne et oracle) en fonction du degré de connaissance sur les positions des formants et des phonèmes. L'initialisation uniforme consiste à choisir un même modèle simple pour tous les phonèmes pour entrainer les HMM. Par contre pour l'initialisation moyenne, chaque phonème a son modèle approprié. L'initialisation oracle correspond à des trajectoires formantiques de référence éditée à la main de 39 phrases prononcées par un locuteur.

L'algorithme a été évalué par rapport à l'influence du nombre d'itérations d'apprentissage ainsi que la dépendance ou l'indépendance des HMM vis-à-vis du contexte phonémique. Cela a permis de constater l'influence de la prise en compte du contexte phonémique sur les performances du suivi, surtout pour le formant F3. Par ailleurs l'impact de l'ordre de la LPC au moment de l'initialisation a été étudié pour trouver l'ordre le plus approprié en fonction du phonème c'est-à-dire le nombre de formants exigé. Il apparaît qu'il est difficile de choisir l'ordre de la LPC de manière sûre.

Il existe d'autres approches de suivi de formants basées sur les GMM tel que l'algorithme proposé par (Darch.J 2005). Cette méthode peut encore être considérée comme une méthode de suivi de formants basée sur le modèle HMM. Nous allons la présenter tout de suite.

- **Approche proposée par (Darch.J 2005)**

Cette méthode proposée par Darch vise à prédire les fréquences de formants à partir d'un flux de vecteurs MFCC (Mel Frequency Coefficients cepstral). La prédiction est basée sur la modélisation conjointe de la densité des vecteurs MFCC et des fréquences des formants données par l'analyse LPC en utilisant l'approche GMM (Gaussien Mixture Model).

Les vecteurs MFCC sont conçus spécifiquement pour la reconnaissance de la parole et constituent une représentation de la parole relativement robuste. Il est en particulier possible de réduire les effets des influences extérieures telles que le bruit à l'aide de différentes techniques. On peut donc attendre aussi une meilleure robustesse pour le suivi de formants.

En revanche la procédure de calcul des vecteurs MFCC entraîne une perte d'informations liée à l'intégration fréquentielle en haute fréquence.

. Les auteurs se sont basés sur la méthode d'extraction des vecteurs MFCC selon la norme de « ETSI Aurora ». (Voir Fig.2.13)

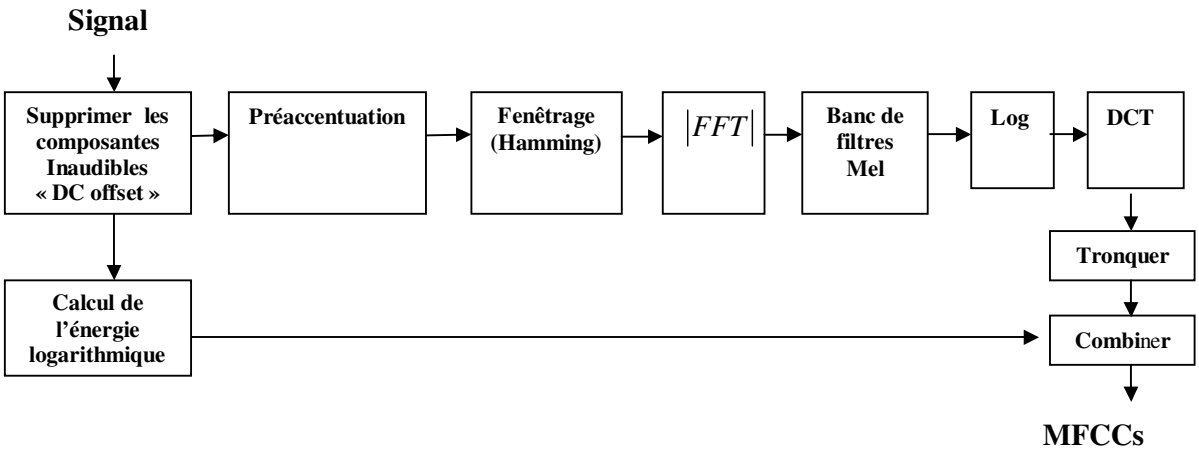


Figure 2. 13 : *Schéma détaillé de l'extraction des vecteurs MFCC selon la norme standard de* « ETSI Aurora » (Darch.J 2005)

L'information de phase est perdue au cours de l'opération du calcul de l'amplitude, alors que les détails spectraux sont perdus durant le filtrage Mel en passant de 128 à 23 canaux et par le biais de la troncature du vecteur cepstral (de 23 au 13 coefficients après l'étape de la transformation de cosinus discrète). Cette perte des détails spectraux cause la dégradation de la précision de l'estimation des formants. Ceci signifie qu'il n'est pas possible d'obtenir des suivis des formants précis en utilisant seulement les vecteurs MFCC.

Toutefois, un travail récent a confirmé qu'on peut prédire le pitch à partir des vecteurs MFCC en modélisant la réunion des densités des MFCC et du pitch en utilisant un GMM. La méthode présentée ici par Darch applique une approche similaire pour la prédiction des fréquences des formants à partir des MFCC.

Les vecteurs MFCC sont extraits en utilisant des trames de durée de 25 ms avec un pas de 10ms. Pour chaque trame de la parole, un vecteur de MFCC contenant 13 coefficients plus d'un terme d'énergie logarithmique est calculé.

Par ailleurs, l'analyse LPC est utilisée pour obtenir une estimation initiale des quatre premiers formants F (Voir Fig.2.14). L'apprentissage du modèle GMM exige la création d'un vecteur y défini comme suit :

$$y_i = [x_i, F_i]^T \qquad \text{(Eq.2.15)}$$

$$\text{où} \quad x = [x_0, x_1, ...x_{12}, \ln(e)] \quad \text{et F=} [F_1, F_2, F_3, F_4]$$

où x est le vecteur MFCC, F désigne le vecteur des quatre premiers formants et i est l'indice où indique le numéro de la trame.

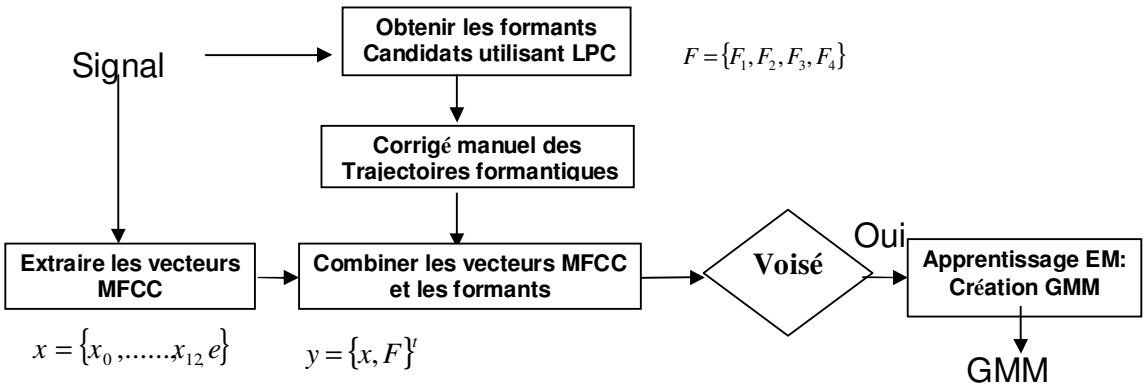


Figure 2. 14 : *La procédure pour obtenir le vecteur final*

Pour s'assurer que les trajectoires des formants initiales estimées par LPC soient exactes, on les corrige à la main si nécessaire. Pour classifier les trames selon le voisement, un simple seuil basé sur l'énergie est utilisé. La classification des zones voisées est par la suite corrigée à la main pour éliminer les parties qui ne contiennent pas de structure formantique. Ensuite, le modèle GMM peut être utilisé pour modéliser les densités des vecteurs MFCC et les formants estimés par LPC. En utilisant les données entraînées associées aux zones voisées, l'algorithme de maximisation de l'espérance (EM) est appliqué pour accomplir un regroupage non supervisé pour ainsi produire le modèle GMM avec K «cluster ». Chaque cluster c_k, modélise la localisation de PDF (Probability Density Funcion) des MFCCs et les formants en calculant la moyenne et la covariance.

L'utilisation de la relation entre MFCC et les fréquences des formants est modélisée par le GMM pour prédire les vecteurs de formants $\hat{F}$. Cette prédiction peut donner soit le cluster le plus proche du vecteur MFCC, soit en utilisant une contribution pondérée des clusters les plus proches. Dans les deux cas le schéma global de la prédiction des vecteurs de formants $\hat{F}$ est présenté par la Fig.2.15.

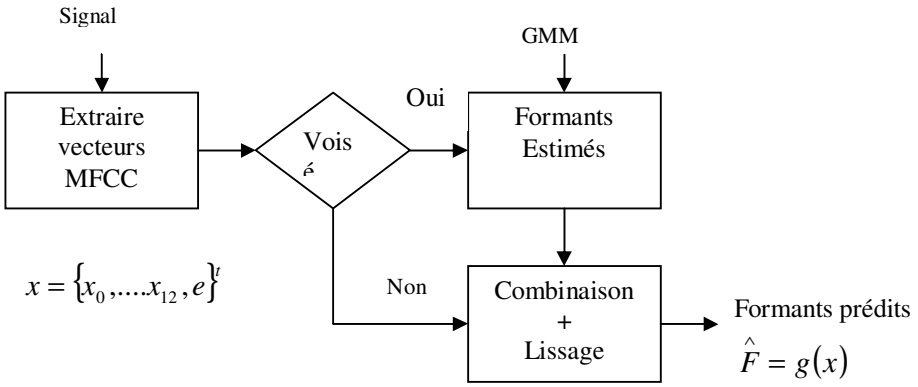


Figure 2. 15 : *Schéma global de la prédiction des vecteurs de formants*

Les deux techniques sont basées sur le critère de maximisation a postériori (MAP).

L'évaluation du suivi repose sur le calcul de pourcentage des fréquences erronées et de l'erreur de la classification des clusters.

D'après les résultats statistiques de l'erreur de classification des clusters et du pourcentage de l'erreur des fréquences des formants, il apparaît que la technique de la prédiction pondérée est plus performante que l'autre.

Aussi les résultats confirment que l'augmentation du nombre des groupes donne une réduction des erreurs de classification des formants ainsi que des erreurs de prédiction.

2.3.3. Approches basées sur le filtre de Kalman

Il y a plusieurs méthodes de suivi qui se basent sur les techniques de filtrage de Kalman. Le principe de ce filtrage adapté au suivi de formant est le fait d'utiliser un système

dynamique de deux équations qui sont une équation d'état (Eq.2.16) et une équation d'observation (Eq.2.17). Ces deux équations sont définies comme suit :

$$x(k+1) = Fx(k) + w(k) \qquad \text{(Eq.2.16)}$$

$$o(k) = Hx(k) + v(k) \qquad \text{(Eq.2.17)}$$

Où x (k) est l'état caché à l'instant k, o(k) est le vecteur d'observation, F est la matrice de prédiction d'états, H est la matrice d'observations, w(k) et v(k) sont des bruits gaussiens non corrélés.

Parmi les techniques populaires de suivi de formants qui ont utilisé le filtrage de Kalman, on présente ici celle de Deng (Deng.L 2004). Cette méthode suit les résonances du conduit vocal (VTR) qui correspondent en fait aux formants surtout pour les voyelles non nasalisées et pour les zones voisées.

Pour assurer la continuité du suivi et pour faire une bonne estimation des VTR, Li Deng a suggéré l'utilisation de modèles dynamiques cachés des VTR pour compenser la probabilité de rater des fréquences candidates pendant le suivi. Pour cela Deng a adopté une forme plus détaillé d'un modèle de parole structuré suivant un système variant dans le temps avec une fonction de prédiction définie par les deux équations d'état et d'observation.

La fonction d'état est définie comme suit :

$$x(k+1) = \Phi x(k) + [I - \Phi]u + w(k) \qquad \text{(Eq.2.18)}$$

Où la fonction (Eq.2.16) est remplacée par la fonction (Eq.2.18). $x(k)$ représente le vecteur d'état caché (définie par : $x = (f, b)' = (f_1, f_2 ... f_p, b_1, b_2 ... b_P)'$). La fonction d'état (Eq.2.18) est dépend du vecteur cible u et de la matrice Φ. Le vecteur caché dynamique est pris égal au VTR formé par les fréquences et largeurs de bande correspondant aux *P* pôles de basse fréquence en utilisant l'analyse LPC. $w(k)$ est un bruit gaussien non corrélé.

La fonction d'observation linéarisée par morceaux est donc définie comme suit :

$$o(k) = A_r . x(k) + d_r + \mu + v(k) \qquad \text{(Eq.2.19)}$$

Où $o(k)$ représente le vecteur d'observation acoustique. A_r est une matrice qui caractérise la transition de la trajectoire pendant le suivi et d_r est une matrice qui caractérise la fin de la trajectoire. Toutes les erreurs dues à l'approximation linéaire par morceaux sont absorbées par le paramètre de prédiction résiduelle noté μ. Il faut noter que la région indexée par r dans (Eq.2.19) est sélectionnée en se basant sur la valeur approximative de VTR notée *x*. La région *r* est déterminée par l'étape de prédiction du filtre de Kalman linéarisé. $v(k)$ est un bruit gaussien non corrélé.

Une fois que le modèle linéarisé défini par les équations (Eq.2.18) et (Eq.2.19) est établi, les algorithmes itératifs de lissage et de filtrage adaptatif de Kalman sont alors appliqués directement pour le suivi des VTR. Pour améliorer le suivi, les auteurs utilisent les nouvelles VTR estimées pour calculer la prédiction résiduelle du cepstre et puis calculer les paramètres qui sont la moyenne et la variance de chaque groupe des VTR et les utiliser pour mettre à jour les paramètres résiduels puis passer à une nouvelle itération de l'algorithme. Les résultats de suivi de cet algorithme semblent être prometteurs, cette efficacité est due à l'utilisation des paramètres dynamiques cachés de la parole. La pertinence de l'algorithme est due aussi à l'application d'un algorithme itératif de filtrage de Kalman adaptatif permettant la linéarisation des composants non linéaires en effectuant le suivi des VTR et à l'entrainement adaptatif des paramètres de prédiction résiduels pour trouver les solutions optimales et améliorer le suivi.

Cependant il y a d'autres approches de suivi de formants plus récentes basées sur le filtrage de Kalman et qui ont apporté des améliorations sur le modèle de filtre telles que les méthodes de Rudoy dans (Rudoy.D 2007) et celles de Özbek dans (Özbek.Y.I 2006) (Özbek.Y.I 2008).

- **Approche proposée par (Rudoy.D 2007)**

Dans cette méthode proposée par (Rudoy.D 2007), les auteurs ont examiné le problème d'estimation des résonances du conduit vocal d'un signal de parole par un modèle gaussien et linéaire.

La plupart des approches paramétriques de suivi de formants visent à estimer les fréquences non observées des VTR et leur largeur de bande. Cependant, la plupart de ces méthodes intègrent des algorithmes supplémentaires (tels que l'analyse LPC pour trouver les racines ou l'extraction des pics du spectre) pour calculer les vecteurs d'observations en fonction des variables cachées ; c'est le fait de formuler une fonction de vraisemblance.

Dans une modélisation d'espace d'état, chaque VTR est modélisé par un résonateur numérique du second ordre et exprimé par une fréquence f_k et une largeur de bande b_k. En supposant que l'enveloppe spectrale de la parole est bien caractérisée par *K* résonateurs, à une trame à l'instant *t* qui peut être modélisée par un vecteur $x_t \in \Re_+^{2K}$ où $x_t = \left(f_1,.....f_K;b_1....b_K\right)^T$, l'évolution des VTR est modélisée par un processus Gaussien-Markovien discrétisé dans le temps défini par :

$$x_{t+1} = Fx_t + w_t \qquad \text{(Eq.2.20)}$$

Où $F \in \Re^{2K*2K}$ est la matrice de transition d'états et w_t est une séquence de bruit blanc satisfaisant l'équation suivante :

$$\mathrm{E}\left(w_j w_j^T\right) = Q\delta_{ij} \qquad \text{(Eq.2.21)}$$

Où Q désigne la covariance de bruit et δ_{ij} désigne le delta de Kronecker définie par :

$$\delta_{ij} = \begin{cases} 0, si(i = j) \\ 1, si(i \neq j) \end{cases} \qquad \text{(Eq.2.22)}$$

A chaque trame à l'instant t, les premiers N coefficients de l'analyse LPC du cepstre sont définis par :

$$y_t = \left(y_t[1],....y_t[N]\right)^T \qquad \text{(Eq.2.23)}$$

La relation entre le vecteur y_t et le vecteur d'état x_t est définie par

$$y_t = h\left(x_t\right) + v_t \qquad \text{(Eq.2.24)}$$

Notons que *h* est le vecteur non linéaire de mappage du vecteur x_t (qui représente les fréquences de résonances et leur largeur de bande) au vecteur y_t des coefficients cepstraux de l'analyse LPC et v_t est la séquence de bruit satisfaisant l'équation suivante :

$$\mathrm{E}\left(v_i v_j^T\right) = R\delta_{ij} \qquad \text{(Eq.2.25)}$$

Où R désigne la covariance de bruit.

Dans cette approche, les auteurs supposent que les bruits w_t et v_t ne sont pas corrélés.

Bien que l'équation (Eq.2.24) soit mathématiquement simple, elle s'applique mal par rapport au filtre de Kalman étendu, dans laquelle une coordination de linéarisation des formants aux coefficients cepstraux désignée par la fonction *h* qui est effectuée par le rapprochement de chaque trame.

Pour cela, selon l'algorithme étendu de filtrage de Kalman proposé par Deng dans (Deng.Li 2006), les auteurs ont proposé dans cet algorithme deux extensions importantes pour réduire l'erreur quadratique moyenne de l'estimation des formants. Premièrement par le fait d'augmenter le vecteur d'état et d'introduire l'estimation de la probabilité en présence des données manquantes, ils obtiennent un modèle qui tient compte explicitement de l'incertitude de la présence de la parole. Deuxièment, ils ont montré comment exploiter la technique de la corrélation croisée entre les trajectoires formantiques pour apporter une amélioration significative à la pertinence de l'estimation des VTR.

En effet, le modèle décrit ci-dessus ne prend pas en compte l'incertitude de la présence de la parole. C'est pourquoi plusieurs approches basées sur ce modèle souffrent de dégradations significatives de leurs performances d'estimation dans la pratique non seulement durant les silences dans les expressions de la parole mais aussi lorsque certains pics des formants ne sont plus observables. Ces effets sont très sensibles dans certaines bases de données telles que la base TIMIT. C'est pour cela que certains chercheurs ont recours à la correction manuelle des erreurs de suivi de formants sur les régions non voisées. Pour éviter ce problème, les auteurs ont appliqué le principe de l'estimation de la probabilité censurée. Il consiste au fait qu'en général, une fonction de probabilité pour un modèle de suivi en présence de données tronquées est formulée à l'aide d'une fonction de vraisemblance qui est la probabilité conditionnelle des données compte tenu des paramètres du modèle.

Les auteurs appliquent ce principe en complétant le vecteur d'état x_t par un indicateur binaire variable pour chaque formant. Les auteurs modélisent ces indicateurs comme statistiquement indépendants d'une trame à une autre et supposent que dans chaque trame ils sont estimés par des détecteurs d'activité vocale. Dans ce modèle d'espace d'état augmenté, l'estimation de l'état optimal peut être obtenue par le filtrage de Kalman classique.

La deuxième extension consiste à modéliser les formants par la technique de corrélation croisée « cross-correlation ». C'est une technique standard destinée à l'estimation du degré de corrélation de deux séries issues d'un seul signal. A titre préliminaire de la phase d'analyse, les auteurs ont estimé empiriquement la corrélation entre les trois trajectoires formantiques corrigées à la main dans la base de données de Deng (Deng.Li 2006). Ils ont appliqué la fonction de corrélation croisée de telle manière qu'un ensemble de valeurs des formants à la trame t puisse être utile pour prédire les valeurs de tous les formants à la trame t+1. Cette observation suggère une autre extension du modèle d'espace d'état décrit ci-dessus. La matrice de transition d'état notée *F* peut être utilisée pour incorporer la structure de la

fonction de corrélation croisée des formants à un décalage d'une trame ce qui permet des améliorations possibles dans l'estimation des paramètres des VTR. Les éléments de la matrice *F* peuvent être estimés en utilisant l'estimation linéaire des moindres carrées. Pour évaluer l'algorithme décrit ci-dessus, les auteurs ont effectué des tests sur la base de données corrigée à la main proposée par Deng. Cette base consiste à un ensemble de phrases de la base TIMIT. Elle contient les vecteurs des quatre premiers formants et leur largeur de bande pour chaque trame traitée.

Pour comparer les résultats avec ceux de Wavesurfer, les auteurs ont fait des tests sur 516 phrases. La phase expérimentale de l'algorithme peut être décrite en trois parties.

-Tout d'abord les auteurs ont appliqué une analyse LPC d'ordre 12 avec une préaccentuation du signal. La fenêtre d'analyse utilisée est Hamming de taille de 20ms et avec un recouvrement de 50%.

-Ensuite, les auteurs ont appliqué un ré-échantillonnage des signaux de parole de 16KHz à 7 KHz dont le but de ne suivre que les trois premiers formants. Puis ils ont calculé les 15 premiers coefficients cepstraux de l'analyse LPC en utilisant la méthode standard décrite plus haut.

Pour chaque expression de parole, un détecteur d'énergie est utilisé pour identifier les trames qui ne contiennent pas de parole (régions non voisées).

-Finalement, les auteurs ont appliqué leur méthode d'ajustement des paramètres du modèle décrite précédemment. Ils ont utilisé la sortie de l'outil Wavesurfer comme un moyen d'estimation empirique des paramètres du modèle. Pour chaque expression, la matrice de transition d'état *F* est estimée à partir des suivis de formants effectués par Wavesurfer. La matrice de covariance du processus bruit de l'équation état notée *Q* est obtenue par le calcul de la probabilité maximale d'estimation conditionnée sur la présence de bruit. Les auteurs ont linéarisé l'équation de mesure (Eq.2.24) en utilisant la fonction de prédiction linéarisée proposée par Deng dans (Deng.Li 2004) pour obtenir la matrice d'observation notée *H*. Enfin, ils ont fixé la matrice d'observation de la covariance de bruit notée *R* à une matrice diagonale pour toutes les expressions. Pour le calcul des trajectoires des formants, ils ont utilisé la version étendue du filtrage de Kalman proposée par Deng dans (Deng.Li 206) pour suivre proprement les formants durant les intervalles de silence. Ils ont aussi implémenté un algorithme de lissage nommé « Rauch-Tung-Stribel » afin de lisser les trajectoires des formants.

Les résultats de l'évaluation indiquent la pertinence de suivi de formants de l'algorithme proposé par Rydoy par rapport à celui donné par Wavesurfer. En revanche, tout

en restant dans certains cas meilleur que Wavesurfer, l'algorithme de Rudoy ne couvre pas les trajectoires entières des VTR de référence particulèrement au niveau des F2 et F3 lorsque ces formants présentent des transitions rapides entre phonèmes. Les résultats de calcul de l'erreur quadratique moyenne de l'algorithme de filtrage de Kalman par rapport à la référence montrent une amélioration marquée due à l'adjonction des deux techniques de l'estimation de la probabilité censurée et le calcul de la corrélation croisée.

- **Approche proposée par (Özbek.I.Y 2006)**

La méthode de suivi de formants proposée par Özbek a apporté une amélioration sur le filtrage de Kalman en utilisent un automate d'états flous pour les différentes décisions sur le début et la fin du suivi.

Les auteurs considèrent que les renforcements spectraux sont des formants quels que soient les sons considérés. Il s'agit d'une extension du terme « formant ».

Les auteurs parlent de Résonances visibles du conduit vocal, VVTR, puisque les résonances du conduit vocal peuvent ne pas être continues en particulier au niveau des nasales où le conduit vocal change structurellement. Les VVTR englobent les formants définies par les sons vocaliques, les formants dus à la nasalisation et les formants dans les régions obstructives. Pour cette raison, la méthode proposée par (Özbek.I.Y 2006) est différente des études précédentes. Le nombre des fréquences de résonances visibles dépendant d'une expression de parole donnée est à priori n'est pas connu et il peut changer dans le temps, c'est ainsi que l'algorithme de suivi doit avoir la capacité de suivre un nombre de fréquences de résonance variable. L'algorithme de suivi doit avoir aussi l'aptitude d'initialiser de nouvelles trajectoires et de terminer certaines trajectoires déjà existantes.

Dans cette étude proposée, Özbek présente une nouvelle stratégie de suivi des trajectoires des VVTR complètement automatique et sans utilisation d'une information phonémique quelconque. L'algorithme de suivi des VVTR est principalement basé sur les algorithmes de suivi « multi-cibles » qui sont largement utilisés dans la littérature. Il fonctionne sur quatre étapes :

- Analyse du signal de parole,
- Décisions du début/fin du suivi,
- Phase d'association des données,
- Suivi en utilisant le filtrage de Kalman.

Les différentes étapes de l'algorithme de suivi sont décrites dans la figure ci-dessous. (voir Fig.2.16)

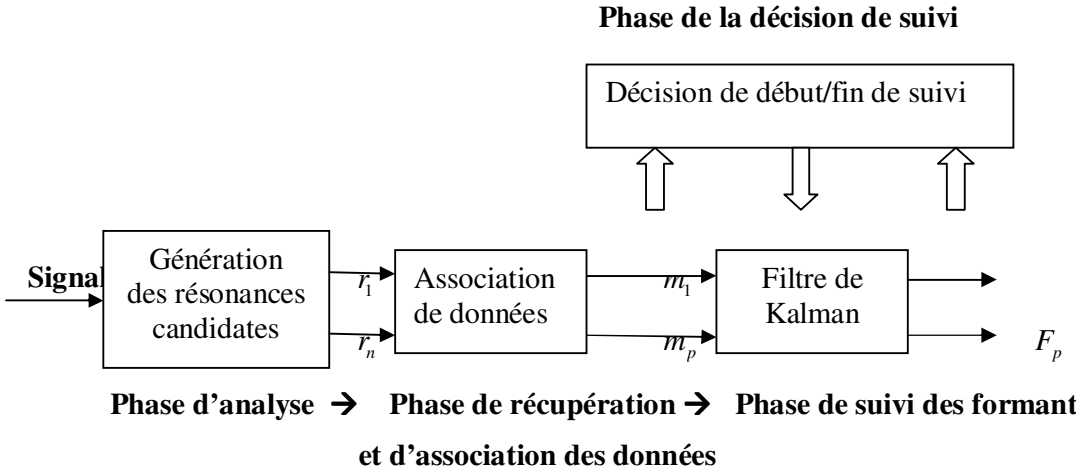


Figure 2. 16 : *Schéma global de l'algorithme de suivi de formants proposé par* (Özbek.Y.I 2006)

Dans la phase de l'analyse du signal de parole, les fréquences de résonance candidates sont obtenues en résolvant le polynôme de la prédiction linéaire LPC. Cependant seuls les pôles complexes sont retenus. Chaque pôle est exprimé par une paire (fréquence de résonance, largeur de bande$\{F_k, B_k\}$). Les largeurs de bande des fréquences de résonance sont utilisées comme seuils pour le processus de sélection des résonances candidates. De nouvelles trajectoires sont crées s'il y a des candidats valables et inversement des trajectoires sont arrêtées s'il n'y a plus de candidats. Cette phase fournit à l'algorithme de suivi trois décisions concernant les trajectoires : commencer une nouvelle trajectoire de VVTR, confirmer ou terminer une trajectoire existante.

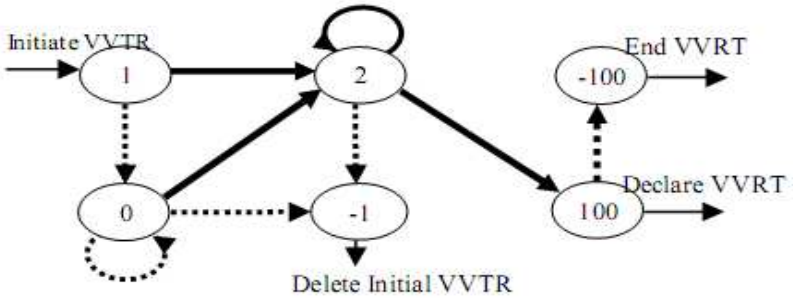


Figure 2. 17 : *Schéma expliquant les différents stades de la phase de décisions début/fin des trajectoires VVTR*

L'état 1 correspond à l'initialisation du suivi. Tout point obtenu comme une fréquence de résonance par l'analyse LPC et qui n'appartient à aucun suivi existant est considéré comme une fréquence candidate d'une nouvelle trajectoire (état 1). A la trame suivante, l'état du suivi candidat passera à l'état 0 ou l'état 2. S'il y a une fréquence de résonance candidate qui est compatible avec le suivi candidat généré à l'état 1, le stade de ce suivi passe à l'état 2 sinon il revient à l'état 0. Les états 0 et 2 sont des états d'attente et les temps d'attente sont les paramètres réglables. A l'état 2 les filtres de Kalman sont initialisés par les candidats. Si le suivi n'a pas reçu de candidats, il passe à l'état 0. Si le suivi ne peut plus accepter de fréquences candidates, il passe à l'état -1 et la trajectoire sera supprimée. Mais dans le cas des fréquences candidates compatibles, il passe à l'état 2. Le suivi passe alors de l'état 2 à l'état 100 et devient donc une trajectoire tant qu'il reçoit des candidats compatibles. L'état 100 est l'état de suivi. Si les trajectoires ne reçoivent plus de candidats pendant un certain intervalle de temps, elles passent alors à l'état -100.

La phase d'association est principalement liée à la phase de décision. Elle détermine les limites supérieures et inférieures des trajectoires des VVTR. Cette phase définit les limites supérieures et inférieures pour les fréquences de résonances candidates. Cette barrière peut être une valeur constante ou variable. La barrière constante est utilisée pour les trajectoires que leur état soit 1, 0 ou 2. La barrière variable est utilisée pour les trajectoires à l'état 100 après que la trajectoire a été initialisée. La barrière change au cours de temps et ses limites supérieures et inférieures sont définies comme suit :

$$Limits_k = \hat{y}_k \mp \sqrt{Th.S_k} \qquad \text{(Eq.2.26)}$$

où $\hat{y}_k$ est le candidat estimé à l'instant *k* qui est obtenu en utilisant l'étape de Kalman et S_k et Th sont respectivement la matrice de covariance et le seuil prédéfini. Cette expression indique que la valeur de barrière change selon la mesure de la covariance. Si une ou plus d'une valeur manque pour une trajectoire quelconque, le voisin le plus proche est intégré à la trajectoire.

Finalement, le filtrage de Kalman est appliqué pour suivre les fréquences de résonance choisies par les étapes précédentes. La représentation du modèle du système dynamique filtre de Kalman est donnée comme suit :

$$\begin{aligned} x_k &= Ax_{k-1} + Gw_{k-1} \\ y_k &= Hx_k + v_k \end{aligned} \qquad \text{(Eq.2.27)}$$

Le système est constitué par un vecteur d'état noté x_k et un vecteur d'observation noté y_k. Selon la théorie proposée par les auteurs, ils supposent que les trajectoires des fréquences de résonance peuvent être modélisées approximativement comme des fonctions linéaires du temps dans des intervalles courts. Les changements non linéaires de la trajectoire (changement brusque dans les trajectoires) sont modélisés via un processus de bruit du système dynamique représenté ci-dessus. Donc les matrices *A*, *G* et *H* sont définies comme suit :

$$A=\begin{pmatrix}1 & T\\ 0 & 1\end{pmatrix}, G=\begin{pmatrix}T^2/2\\ T\end{pmatrix}, H=[1,0] \qquad \text{(Eq.2.28)}$$

Où *T* est l'intervalle de mesure qui est constant.

Le vecteur d'état x_k est défini comme suit :

$$\mathbf{x_k}=\begin{bmatrix}F_k & \dot{F}_k\end{bmatrix}^T \qquad \text{(Eq.2.29)}$$

Où F_k est la fréquence de résonance correspondante qui est suivie et $\dot{F}_k$ est sa dérivée temporelle. y_k est le vecteur de mesure obtenu par les résonances candidates. w_k et v_k sont définies comme des mesures de bruit blanc gaussien. L'algorithme de suivi utilise les équations de filtrage de Kalman où les mesures sont les fréquences obtenues à la phase d'association des données.

Pour évaluer l'algorithme, les auteurs ont effectué des tests sur des expressions de la parole continue prise d'une base de données turque. Ils ont utilisé l'outil (Wavesurfer) pour comparer les résultats des VVTR trouvées par l'algorithme proposé. Il apparaît clairement que cet algorithme assure bien la continuité du suivi. Par ailleurs, chaque fois que le conduit vocal change fortement ce qui traduit par un changement brusque de ses fréquences de résonance, l'algorithme trouve des trajectoires qui s'arrêtent et d'autres qui commencent. Les résultats ont montré que les VVTR trouvées par l'algorithme proposé par Özbek, donne des résultats plus satisfaisants que Wavesurfer surtout au niveau des consonnes vélaires, des nasales /m/, de la voyelle /e/ et même au niveau de la voyelle /a/ lorsqu'elle suit la liquide /l/.

Peu d'années après, Özbek a essayé dans des travaux plus récents (Özbek.Y.I 2008) d'améliorer son algorithme de suivi de formant basé sur le filtrage de Kalman en ajoutant comme extension l'utilisation de la programmation dynamique. Dans ce travail, le signal est segmenté en des parties voisées et non voisées, les résonances du conduit vocal sont estimées

en utilisant la programmation dynamique et en plus elles sont traitées par le filtrage et le lissage de Kalman pour lisser la trajectoire de chaque formant.

- **Approche proposée par (Özbek.Y.I 2008)**

Dans cette approche, Özbek (Özbek.Y.I 2008) a combiné le filtrage de Kalman avec l'algorithme de la programmation dynamique pour estimer et suivre les fréquences des formants avec précision. En faisant cette combinaison, on considère le processus de suivi de formants comme une sorte de processus de suivi multi-cibles (multi-formants). Dans les applications de suivi multi-cibles, il y a deux décisions importantes à prendre qui sont l'association des données (quelle mesure appartient à quelle cible ?) et l'estimation de leurs positions. En s'appuyant sur cette idée, l'auteur a considéré les formants candidats prédits par l'analyse LPC en tant que mesures cibles correspondant à des fréquences de formants. La programmation dynamique est considérée comme étant l'étape d'association des données dans laquelle l'étiquetage des formants candidats est traité. L'estimation des positions des formants est faite dans l'étape de filtrage de Kalman. La segmentation du signal en des parties voisées et non voisées est l'un des facteurs qui améliore les performances du système. Pour la parole voisée comme pour la parole non voisée, les fréquences nominales des formants sont utilisées comme une information complémentaire dans l'algorithme de programmation dynamique pour assurer la continuité du suivi.

Voici le fonctionnement détaillé de l'algorithme. Le signal est d'abord segmenté à l'aide de l'algorithme de programmation dynamique LBDP (Level Building Dynamic Programming) proposé par (Sharma.M 1996). Après la phase de segmentation, chaque segment est classé comme voisé ou non voisé en utilisant deux seuils (énergie moyenne en dB du segment traité et le taux d'énergie en dB de la bande basse fréquence de 100-900 Hz et de la bande haute fréquence 3700-5000Hz). Le signal est préaccentué et décomposé en trames de 40 ms. Pour chaque trame, les fréquences et les largeurs de bandes des formants candidats sont calculées à partir du polynôme de la fonction de transfert donné par l'analyse LPC (ordre 12). Une fois que les fréquences candidates des formants sont prédites, ces données ont été étiquetées en formants en utilisant la programmation dynamique pour associer les résonances candidates avec la plage de fréquence de chaque formant. La programmation dynamique utilisé un coût local et un coût de transition. Le coût local est relié aux connaissances sur les valeurs nominales des formants sans utiliser aucun contexte temporel. La fonction du coût

local contient le paramètre de la largeur de bande et la différence moyenne normalisée entre les valeurs candidates et nominales des fréquences de résonances. Le coût de transition permet d'imposer une contrainte de continuité. Contrairement au système classique nommé Baseline proposé par (Talkin.D 1987), (Xia.K 2000), pour la méthode proposée ici est une version étendue du système Baseline car les paramètres de la programmation dynamique sont fixés différemment pour les parties voisées ou non voisées.

En fait, le système Baseline est le modèle classique de suivi de formants basé sur l'analyse LPC et la programmation dynamique mais sans segmenter le signal en parties voisées et non voisées et sans utiliser le filtrage de Kalman.

L'estimation finale des VTR, Özbek fait appel à la technique standard de filtrage et de lissage de Kalman. La méthode proposée ici a été comparée avec le système Baseline, la méthode proposée par (Deng.L 2004) et l'outil public (Wavesurfer).

Les expériences réalisées en utilisant la base de données étiquetée à la main par Deng. Li (Deng.Li 2006). Les erreurs d'estimation des formants sont calculées pour les grandes classes phonétiques. Les résultats expérimentaux montrent que la méthode proposée est significativement meilleure que le système Baseline et Wavesurfer et plus performante que la méthode de Deng spécialement pour les classes des voyelles et semi-voyelles. Mais les performances de cette méthode sont faibles pour les nasales à cause de l'analyse LPC qui ne trouve pas les pics pertinents.

2.3.4. Approches basées sur les courbes

- **Méthode proposée par (Sun.X 1995)**

  Cette technique consiste à estimer les trajectoires des formants contenant l'énergie dense du spectrogramme. Ces trajectoires représentent l'ensemble des centres de gravité pour chaque bande d'énergie. Chaque centre de gravité est calculé à partir d'un ensemble de pics spectraux pour chaque intervalle de temps. Pour cela, l'auteur utilise une fonction spline cubique pour l'estimation des centres de gravité qui dépend de deux paramètres : un pour le lissage des courbes et l'autre pour que la valeur estimée soit la plus proche possible des valeurs d'observations. L'entrée de cette fonction d'estimation peut être initialisée par les valeurs des pics du spectre de LPC. L'ajustement de ces trajectoires se base sur la méthode des moindres carrés en

utilisant un mélange de modèles de spline multiple en appliquant un algorithme itératif puisque dans le cas traité, l'auteur est en train de suivre dans le spectrogramme les trois trajectoires correspondantes aux trois premiers formants. Cette approche proposée par Sun présente deux avantages majeurs. Le premier avantage est le fait qu'elle soit robuste contre les pics manquants et les faux pics qui auraient lieu. Le deuxième avantage est le lissage des trajectoires qui est garanti grâce aux propriétés de régression propre aux modèles spline.

- **Méthode proposée par (Yves.L 2004)**

  C'est une méthode basée sur la technique de « true enveloppe » et celle des courbes concurrentes. Elle consiste à élaborer un algorithme de suivi itératif qui déforme les courbes représentant les formants sous l'influence de l'énergie du spectrogramme. C'est le concept de l'approche variationnelle des contours actifs « snake ». Récemment l'auteur a ajouté une stratégie de contrôle des formants qui consiste à utiliser une force de répulsion entre les formants pour qu'ils partagent l'énergie du spectrogramme de manière optimale.

2.4. Conclusion

Nous avons présenté dans ce chapitre tout d'abord, des techniques d'estimation des fréquences des formants basées principalement sur l'analyse de la prédiction linéaire, le lissage cepstral et les modèles auditifs. Ensuite, nous avons présenté certaines techniques de suivi de formants telles que celles basées sur la programmation dynamique, d'autres basées sur les modèles HMM et nous avons terminé par présenter quelques approches de suivi basées sur le filtrage de Kalman. Cette analyse bien détaillée de quelques méthodes d'estimation et de suivi de formants va nous permettre de mieux comprendre les difficultés du suivi et de proposer une nouvelle technique de suivi de formants.

Chapitre III
Préparation de la Base de Données

3.1. Introduction

Malgré l'importance du rôle joué par les fréquences des formants pour la perception et le traitement de la parole, on constate qu'il n'y a peu voire pas de bases de données de référence, surtout en langue arabe. Comme ces bases sont nécessaires à l'évaluation quantitative des techniques automatiques de suivi des formants, nous présentons dans ce rapport nos efforts récents visant à la préparation d'un corpus de référence en langue arabe étiqueté phonétiquement et avec les trajectoires des trois premiers formants (Rekhis.O 2009).

Dans ce chapitre nous allons tout d'abord décrire quelques aspects linguistiques et phonétiques de la langue arabe standard, ensuite, présenter brièvement notre corpus ainsi que le logiciel Winsnoori (Wins) avec lequel nous avons fait l'étiquetage. Enfin, on va terminer par une description détaillée de l'étiquetage manuel phonétique et formantique ainsi que les difficultés rencontrées lors de cet étiquetage.

3.2. Description phonétique et linguistique de la langue Arabe standard

Par opposition à l'arabe classique ancien (langue de la poésie préislamique), à l'arabe coranique (langue du Coran), et à l'arabe classique post-coranique (langue de la civilisation arabo-musulmane), on distingue l'arabe standard moderne qui naît au début du XIX^{e} siècle en

Égypte, qui est enseigné dans les écoles contemporaines et parlée dans le cadre officiel. C'est la langue écrite commune de tous les pays arabophones.

Dans cette section, nous présentons une étude exhaustive des caractéristiques phonétique, phonologique et syllabique de la langue arabe. Nous décrivons les systèmes vocaliques et consonantiques et leurs classifications selon le mode d'articulation ainsi qu'une description brève de quelques aspects phonologiques de cette langue. Par la suite, nous présentons les principales syllabes de la langue arabe et leurs différentes caractéristiques.

3.2.1. Caractéristiques de la phonétique et la phonologie de la langue Arabe

L'alphabet phonétique arabe comporte 28 consonnes, 6 voyelles et quelques autres réalisations vocaliques (Saidane.T 2004). Les phonèmes arabes se distinguent par la présence de deux classes qui sont appelées pharyngales et emphatiques (Satori.H 2007).

En ce qui concerne les voyelles de l'arabe, on distingue trois voyelles courtes et trois voyelles longues. La durée d'une voyelle longue est environ le double de celle d'une voyelle courte. Elles sont représentées dans les Tableaux 3.1 et 3.2 ci-dessous:

Courtes	Longues
ُ/ ِ/ ◌ْ/	ا ـِي ـُو

Tableau 3. 1 : *Graphèmes des voyelles en langue arabe*

æ ◌َ /ɑ/	i ◌ِ /i/	u ◌ُ /u/	ææ ا◌َ /ɑɑ/	i i ◌ِي /i i/	uu ◌ُو /uu/

Tableau 3. 2 : *Notation internationale des voyelles en langues arabes*

Les trois voyelles principales se trouvent aux extrémités du triangle vocalique (Boukadida.F 2006) : /a/, /i/ et /u/ (appelées respectivement *fatha*, *kasra* et *dhamma*). Elles sont caractérisées par deux classes de localisation : antérieure étirée (/i/), postérieure arrondie (/u/) et par deux degrés d'aperture : fermé (/i/ et /u/) et ouvert (/a/). Dans la littérature, certains considèrent qu'on a aussi une quatrième voyelle : /°/ (appelée soukoun), c'est une voyelle muette.

Il faut noter tout d'abord que les voyelles courtes sont facultatives, on écrit généralement les textes sans ces voyelles.

On les appelle en arabe « El Haraket » et elles sont écrites de la manière suivante

-la voyelle /a/ s'écrit au dessus de la lettre,

-la voyelle /u/ s'écrit au dessus de la lettre,

-la voyelle /i/ s'écrit au dessous de la lettre,

-la voyelle muette /°/ s'écrit au dessus de la lettre.

Ces lettres bien entendues sont des consonnes.

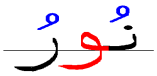
Un exemple de phrase avec les voyelles courtes (Arabic 2011):

KATABA OMAR A
'OUMAROU ECRIT

La longueur de la voyelle diffère selon la position de cette dernière dans le mot et selon le nombre de syllabes de ce mot (Braham.A 1997). La longueur de la voyelle en arabe est contrastive; une voyelle courte et sa contrepartie longue peuvent donner lieu à deux mots différents (Zaki.A 2004).

Les voyelles longues sont : « alif »/ ا/, « waw » /و/ et « ya »/ ى/. Exemples de voyelle longue (Arabic 2011):

 M A DH A QUOI

 N OU R OU LUMIERE

 ' I D I MA FETE

Remarque :

Ces trois lettres sont à la fois des voyelles et des consonnes (semi-voyelles) (Arabic 2011), elles sont associées aux trois voyelles brèves pour former des voyelles longues appelées en langue arabe « madd ». Lorsque ces trois phonèmes ne jouent pas le rôle de voyelles, ils se comportent comme les autres consonnes, et peuvent donc « porter » des voyelles brèves.

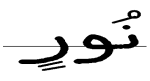
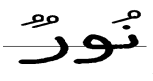
Ces voyelles peuvent avoir des timbres différents selon leur contexte d'apparition. Par exemple en fin de mot, la voyelle peut être brève, et certaines fois (selon la grammaire) doublée ce qu'on appelle « tanween ». C'est à dire de la forme suivante (Arabic 2011):

« Tanween » est la prononciation de la lettre « n » à la fin du mot.

Il y a trois formes du « tanween »:

- « Tanween avec « fatha »: constitué de 2 « fatha » /a/ situées au dessus de la dernière lettre et surtout ne pas oublier d'ajouter « alif ». La prononciation sera donc d'ajouter à la fin du mot AN ».
- « Tanween avec « dhamma » : constitué de 2 « dhamma » /u/ situées au dessus de la dernière lettre. La prononciation sera donc d'ajouter à la fin du mot OUN ».
- « Tanween avec « kasra »: constitue de 2 « kasra » /i/ situées au dessous de la dernière lettre. La prononciation sera donc d'ajouter a la fin du mot IN ».

On écrira par exemple:

 *NOUR AN*  *NOUR IN*  *NOUR OUN*

On remarque aussi que dans un contexte emphatique (au contact des consonnes / sˤ/, / dˤ/, / ðˤ/ et / tˤ/ voir tableau 4.3), le point d'articulation des voyelles est déplacé à l'arrière ce qu'on appelle les voyelles emphatiques (Boukadida.F 2006).

On note également la présence de deux diphtongues, qui sont (/aw/ et /ay/). Les diphtongues se produisent quand les deux semi-voyelles /w/ et /y/ sont précédés par la voyelle /a/ (Zaki.A 2004)**.** Par exemple, le mot /jayyid/ dans lequel /y/ est précédé par la voyelle courte /a/ produit une diphtongue /ay/.

Concernant les consonnes, nous pouvons les classer selon plusieurs critères : leur voisement et leur lieu et mode d'articulation. (Voir Tableau 3.3)

	Labiale	Labio-dentale	Interdentale	Dental/ Alvéolaire	Palatale	Palatale/ Alvéolaire	Vélaire	Uvulaire	Pharyngal	Glottale
Nasale	m			n						
Occlusive Non voisée				t tˤ dˤ			k	q		ʔ
Occlusive Voisée	b			d						
Fricative Non voisée		f	θ	s sˤ		ʃ		x ~ χ	ħ	h
Fricative Voisée			ð ðˤ	z		dʒ ~g^3		ɣ ~ ʁ	ʕ	
Tap				r						
Latérale				l						
Semi-voyelle	w				j		w			

Tableau 3. 3 : *Les articulations des consonnes en arabe selon l'API (Alphabet Phonétique International) (voir Annexe A)*

Le phonème /w/ est associé à deux lieux d'articulation différents. Cela s'explique par le fait qu'il est à la fois articulé avec le rétrécissement de l'ouverture de la lèvre, qui le rend bilabial et l'élévation du dos de la langue vers le palais mou, qui le rend vélaire. Le /w/ en arabe est classifié comme semi-voyelle labio-vélaire (Zaki.A 2004).

La distribution ou la fréquence d'apparition des consonnes en langue arabe diffère d'une consonne à une autre. En effet, la consonne la plus fréquente en langue arabe est /r/ et la plus rare est /ðˤ/.

Les caractéristiques phonologiques de l'arabe sont : la gémination, l'emphase, le « madd », l'indéterminisme « tanween » et l'assimilation (Boukadida.F 2006). Nous venons de présenter le « madd » (voyelle longue) et le « Tanween » (double voyelle).

La gémination « tachdid » d'une consonne est la répétition immédiate de cette même consonne. Au niveau graphique elle est caractérisée par le signe de la « chad-da », c'est un petit signe noté « ّ » qui est a la forme de la lettre "Sin" situé au dessus de la consonne (Boukadida.F 2006). D'autre part, une consonne géminée est un son unique pour lequel les organes de phonation ne changent pas de position. En terme phonétique, cependant, la distinction entre les géminés et les non géminés est faite par rapport à la durée. Au niveau phonologique, il est important de noter qu'en arabe les segments géminés ne peuvent jamais apparaître en début de mot (Newman.D). La gémination détermine un sens et joue un rôle structural dans le développement morphologique nominale et verbal par exemple: /حَضَرَ/ « il a assisté » est différent de /حَضَّرَ/ « il a préparé » (Boukadida.F 2006).

L'emphase se distingue par la postérisation ; c'est-à-dire le recul de la langue vers la partie postérieure de la cavité buccale. Les linguistes arabes ont désigné ces consonnes comme consonnes emphatiques /ص/ / sˤ/, /ض/ / dˤ /, /ط/ / tˤ / et /ظ/ / ðˤ/ (Boukadida.F 2006).

L'assimilation est le transfert d'une caractéristique du trait phonétique d'un son sur un autre immédiatement voisin, ces deux sons tendent alors à devenir semblables au niveau de l'écoute et au niveau de la prononciation. Cette assimilation peut être totale ou partielle (Boukadida.F 2006).

Exemple :

- Au contact d'une vélaire, le /ت/ / t / s'emphatise et devient / ط / / T / : / صوت / ➔ /سوط/

- Dans un certain nombre de conjonctions, de pronoms, il y a l'assimilation de /ن/ /n/ par / م / / m / ou par /ل/ / l / : / أن ما / ➔ / أمّا /, / انْ لَا / ➔ / ألّا /.

3.2.2. La syllabe

La syllabe joue le rôle de l'unité minimale pour la réalisation des traits distinctifs des phonèmes. En effet, un phonème ne peut être prononcé qu'au niveau de la syllabe, les contrastes syllabiques sont nécessaires pour sa classification (Zaki.A 2004).

A propos de la syllabe dans la langue arabe, les chercheurs arabes ont prouvé que l'articulation de certaines syllabes est spécifique au pays d'origine du locuteur puisque chaque locuteur arabe est influencé dans sa prononciation de l'arabe classique par son dialecte habituel spécifique à sa ville natale et son pays (Braham.A 1997). La plupart de temps, il s'agit de petites différences.

La langue arabe comporte cinq types de syllabes (Braham.A 1997) (Satori.H 2007) classées selon les trais ouvert/fermé. Une syllabe est dite ouverte (respectivement fermée) si elle se termine par une voyelle (respectivement une consonne), où le V désigne une voyelle et le C représente une consonne. On peut également classer les syllabes en fonction de leur longueur : on parle de syllabes courtes, de syllabes longues et de syllabes sur-longues (Zaki.A 2004) comme le montre le tableau 3.4 ci-dessous.

Type syllabe	Exemple en arabe	Transcription Phonétique	Traduction En français	Description des syllabes	
CV	بِ	Bi	Avec	Ouverte	Courte
CVV	مَا	Maa	Marqueur d'exclamation ou de négation	Ouverte	Longue
CVC	مِنْ	Min	De	Fermée	Longue
CVVC	بَابْ	Baab	Porte	Fermée	Sur-longue
CVCC	مَهْرْ	Mahr	Poulain	Fermée	Sur-longue

Tableau 3. 4 : *Type de syllabes de la langue arabe standard*

Les quatre premiers types peuvent apparaître dans n'importe quelle position du mot. Le dernier type peut se produire à la fin du mot ou seul comme un seul mot. Les syllabes dans un mot unique ne sont pas prononcées avec le même niveau de volume sonore. Il est possible de trouver trois différents niveaux d'intensité dans le même mot. La cause de ces différences est le stress qui est l'intensité en termes d'énergie du signal de la syllabe. Il y a trois degrés de stress : le stress primaire (principal), le stress secondaire et le stress tertiaire (faible). L'emplacement des différents types de stress dans le mot arabe dépend des types de syllabes, leur distribution et leur nombre (Youssef.A 2004).

Toutes les syllabes commencent par une consonne suivie d'une voyelle, mais celles qui sont le plus utilisées sont : CV, CVC et CVCC. La syllabe CV est la plus fréquente, elle peut se trouver au début, au milieu ou à la fin du mot (Satori.H 2007).

3.3. Présentation du corpus

Pour bien étudier le suivi de formants et faire un bon étiquetage, nous avons besoin de phrases simples et courtes à la suite d'une recherche bibliographique approfondie, nous avons trouvé que les phrases les plus appropriées à notre objectif sont celles préparées par Malika Boudraa (Boudraa.M 2000). En effet, la base de données proposée est constituée de 20 listes formées de dix phrases courtes chacune (voir Annexe B). Elle contient deux types de phrases (déclaratives et interrogatives) dont la plupart sont extraites du coran, du Hadith et de proverbes arabes. La conception de cette base est inspirée de l'approche de l'équilibrage phonétique utilisée par Combescure pour son corpus en langue française (Combescure.P 1981). En premier lieu, le corpus doit couvrir l'ensemble des réalisations phonétiques, phonologiques et phono-tactiques de la langue arabe standard pour être équilibré phonétiquement. Ensuite, le corpus doit répondre à un certain nombre de structures syntaxiques, grammaticales et modales rencontrées lors de la lecture d'un texte arabe.

Chaque liste du corpus est constituée de 104 CV ce qui donne 208 phonèmes. Cette base de données est aussi basée sur les études statistiques de Moussa (Boudraa.M 2000) sur la langue arabe standard en respectant la fréquence d'apparition de chaque phonème.

Nous avons enregistré ce corpus au laboratoire LORIA dans des conditions très favorables. L'acquisition des données a été effectuée sur un PC en utilisant le logiciel « recorder.tcl » sous format .wav (Mono) codée sur 16 bits et à une fréquence d'échantillonnage de 16Khz. L'enregistrement des 20 listes a été fait avec dix jeunes

locuteurs, cinq hommes et cinq femmes, dont les âges varient entre 22 et 30 ans, ce qui a donné au total 2000 enregistrements.

3.4. Description du logiciel Winsnoori

Pour préparer notre base de données, nous avons étiqueté manuellement les phrases avec le logiciel (Winsnoori). Ce dernier est un logiciel d'analyse de la parole. Il fournit cinq types d'outils qui servent à :

- éditer les signaux de parole,
- annoter phonétiquement ou orthographiquement des signaux de parole. Winsnoori offre des outils pour explorer des corpus annotés automatiquement,
- analyser la parole à l'aide de plusieurs analyses spectrales et affichage des pics spectraux au fil du temps,
- étudier la prosodie. Outre le calcul de pitch, il est possible de synthétiser de nouveaux signaux en modifiant la courbe de F0 et / ou le débit de parole,
- générer des paramètres pour le synthétiseur Klatt. Une interface graphique conviviale avec des outils de synthèse par copie (suivi de formants automatique, réglage automatique de l'intensité…) permet à l'utilisateur de générer des fichiers pour le synthétiseur de Klatt facilement.

L'outil Winsnoori peut calculer six types de spectrogrammes et détecter les pics spectraux obtenus par lissage cepstral, l'algorithme de l'enveloppe vraie et le filtre de la prédiction linéaire (LPC). Ces outils peuvent servir à la synthèse automatique (suivi automatique de formants, ajustement des amplitudes des formants…) grâce au synthétiseur de Klatt (Winsnoori).

3.5. Etiquetage phonétique

Pour faire un bon étiquetage formantique qui sera considéré après comme suivi de référence, nous avons d'abord commencé par annoter phonétiquement les signaux du corpus. Le but de la segmentation phonétique de chaque signal de parole est de connaître les instants de début et de fin de chaque son pour pouvoir vérifier sur cet intervalle la valeur fréquentielle des formants avec la valeur nominale correspondant aux valeurs attendues par ce phonème

dans la phrase étudiée. Ainsi, pour préparer notre corpus, nous avons annoté manuellement les signaux du corpus en utilisant Winsnoori.

Nous avons utilisé la transcription selon le code (SAMPA) (Speech Assessement Methods Phonetic Alphabet) pour cet étiquetage phonétique (Voir Tableau 3.5).

Fricatives		Occlusives		Latérale		Trill		Nasale		Semi-voyelles	
Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.
/f/	/ف/	/E/	/ء/	/l/	/ل/	/r/	/ر/	/m/	/م/	/w/	/و/
/s/	/س/	/q/	/ق/					/n/	/ن/	/j/	/ي/
/s./	/ص/	/d./	/ض/								
/z/	/ز/	/d/	/د/								
/h/	/ه/	/b/	/ب/								
/X/	/خ/	/k/	/ك/								
/S/	/ش/	/t./	/ط/								
/T/	/ث/	/t/	/ت/								
/x/	/خ/										
/D/	/ذ/										
/z./	/ظ/										
/G/	/غ/										
/H/	/ع/										
/Z/	/ج/										

Tableau 3. 5 : *Transcription graphème-phonème de la langue arabe standard utilisant le code SAMPA*

Pour l'arabe en notation SAMPA, il y a six voyelles simples et six voyelles colorées à cause de leur timbre qui diffère au contact de certaines consonnes emphatiques (SAMPA) :

- 3 voyelles courtes : /a/, /u/, /i/
- 3 voyelles coutres colorées : /a. /, /i. /, /u. / (après /d. /z. /t. /s./)
- 3 voyelles longues: /A/, /U/, /I/
- 3 voyelles longues colorées : /A. /U. /I. / (après /d. /z. /t. /s./)

Winsnoori est un logiciel qui n'est pas conçu pour la langue arabe, nous avons donc créé un fichier « arabic.pho » que nous avons intégré à ce logiciel pour pouvoir effectuer l'étiquetage (Rekhis.O 2009). Ce fichier contient les différentes classes des phonèmes de la langue arabe accompagné chacun par un exemple de mot contenant ce phonème. A l'aide de ce fichier, nous avons fait l'annotation phonétique manuellement. Pour chaque signal, un

fichier.phn qui contient l'alignement phonétique/temporel de chaque phrase c'est-à-dire la durée temporelle (début/fin en ms) et l'étiquette de chaque son de la phrase traitée a donc été crée.

Exemple : Voici un exemple de l'étiquetage phonétique manuel sur le spectrogramme de Winsnoori ainsi que le « fichier.phn » correspondant à la phrase : «أَيْنَ المُسَافِرُونَ ؟» « **3ayna lmusa:firu:na ?** » (Où sont les voyageurs ?) prononcée par un locuteur.

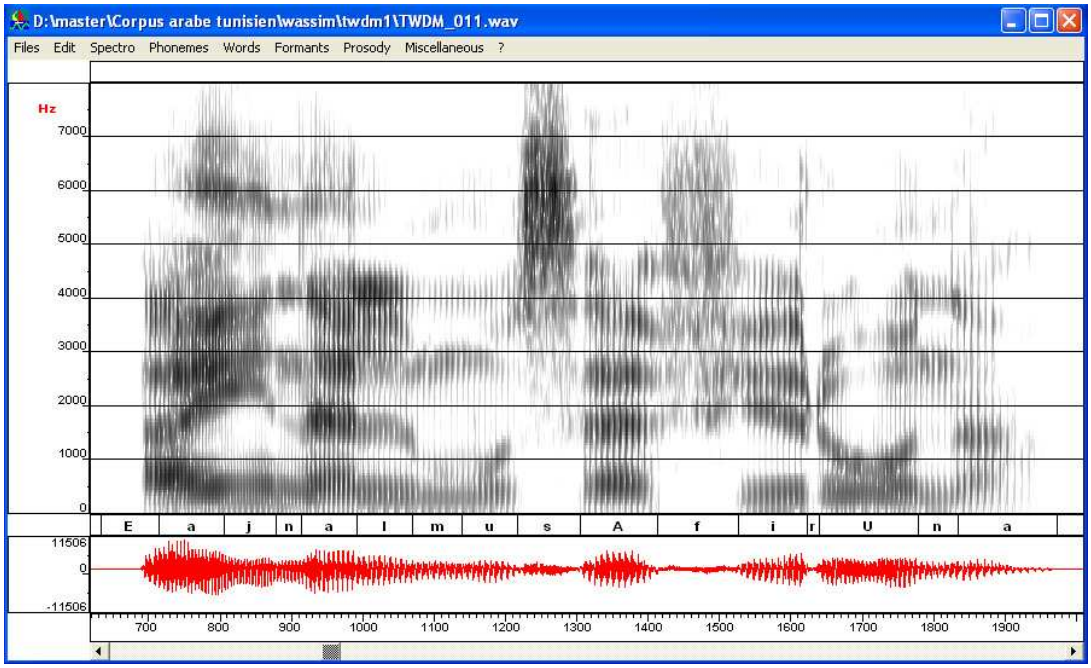


Figure 3. 1 : *Spectrogramme de l'enregistrement* «أَيْنَ المُسَافِرُونَ ؟» « **3ayna lmusa:firu:na ?** » *prononcé par le locuteur Loc.M.4*

```
.ali
freq_echt 0
                693          743    E
                743          805    a
                805          878    j
                878          915    n
                915          991    a
                991         1071    l
               1071         1138    m
               1138         1216    u
               1216         1306    s
               1306         1413    A
               1413         1526    f
               1526         1624    i
               1624         1640    r
               1640         1778    U
               1778         1834    n
               1834         1973    a
               1973            0    FIN

E:\master\Corpus arabe tunisien\wassim\twdm1\TWDM_011.wav
```

Figure 3. 2 : *Ficher.phn de l'enregistrement* «أَيْنَ المُسَافِرُونَ ؟» « **3ayna lmusa:firu:na ?** » *prononcé par le locuteur Loc.M.4*

A ce stade de travail, nous avons rencontré plusieurs difficultés spécialement pour les consonnes. Nous avons donc fait l'appel à des experts en phonétique. Voici énumérées les difficultés les plus importantes (Rekhis.O 2009) :

- Deux consonnes occlusives non voisées /t/ et /E/ se suivent parce qu'il y a une voyelle muette /°/ au dessus de la consonne /t/ et en plus elle n'a pas d'explosion. Pour surmonter ce problème on a eu recourt au signal temporel. (Voir Fig.3.3)

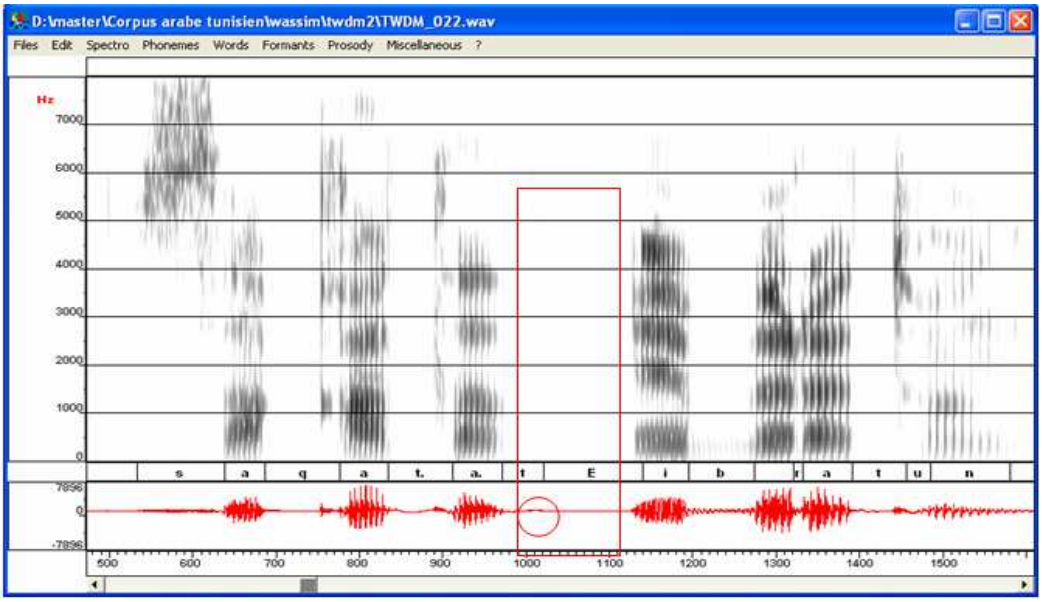


Figure 3. 3 : *Spectrogramme de l'enregistrement* «سَقَطَتْ إِبْرَةٌ.» « **Saqatat 3ibratun** » (Une aiguille est tombée) *prononcé par le locuteur Loc.M.4*

- La présence d'une occlusive non voisée « E » entre 2 voyelles apparaît comme une fricative voisée due au phénomène de coarticulation. Les limites de ce son seront définies à l'aide du « gating » lors de l'écoute (Voir Fig.3.4).

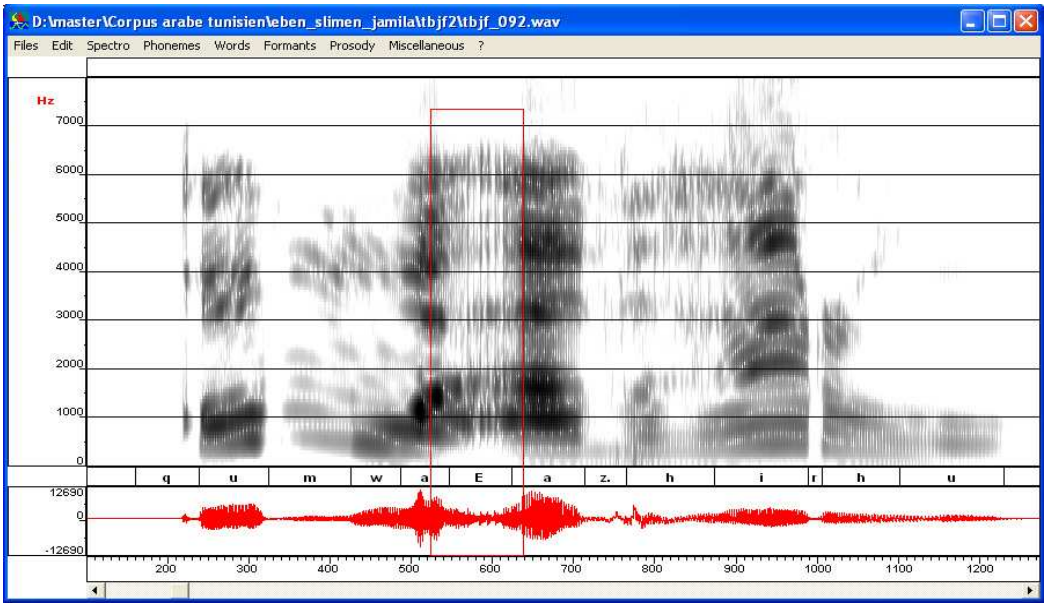


Figure 3. 4 : *Spectrogramme de l'enregistrement* « قُمْ وَأَظْهِرْهُ » « **Qum wa 3azhirhu** » (Vas'y montre le !) *prononcé par la locutrice Loc .W.2*

Les fricatives voisées possèdent des pseudo-formants dont elles prennent les formants des voyelles précédentes et suivantes. L'énergie des phonèmes « h » et « H » est très condensée au niveau du spectrogramme ce qui rend difficile leur localisation au sein de la syllabe VCV.

- Dans certains enregistrements il y a un problème d'allongement final. Le dernier phonème attendu est soit inexistant, soit très court, soit très long et cela dépend du locuteur auquel nous avons fait appel.

- Selon le linguiste Mr. Ghazali, le locuteur tunisien ne fait pas de différence au niveau de prononciation entre l'occlusive non voisée « d. » et la fricative voisée « z. ». Pour cela, il est convenu de prendre toujours « z. » au lieu de « d. ». (Voir Fig.3.5).

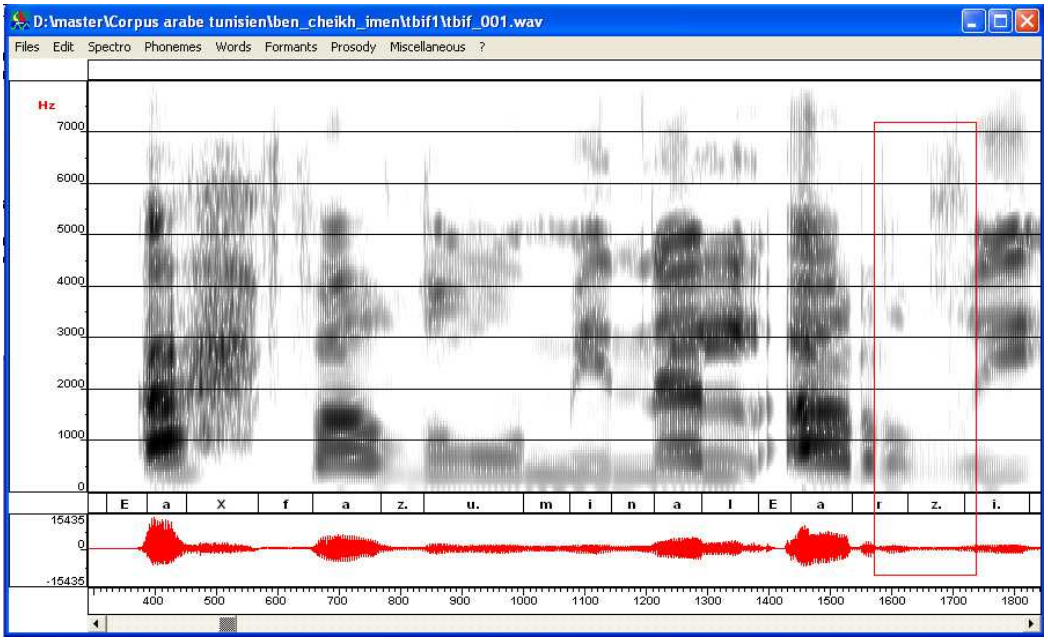


Figure 3. 5 : *Spectrogramme de l'enregistrement* «أَحْفَظُ مِنَ الأَرضِ» « **ɜahfazu mina lɜardi** » (J'apprends de la terre) *prononcé par la locutrice Loc.W.3*

- Souvent lorsqu'une consonne occlusive non voisée est précédée d'une voyelle il y a un bruit qu'on appelle coup de glotte qui est dû au phénomène de coarticulation. Ce bruit apparaît lors de la fermeture de l'occlusive et n'est pas synchronisé avec la voyelle (Voir Fig.3.6).

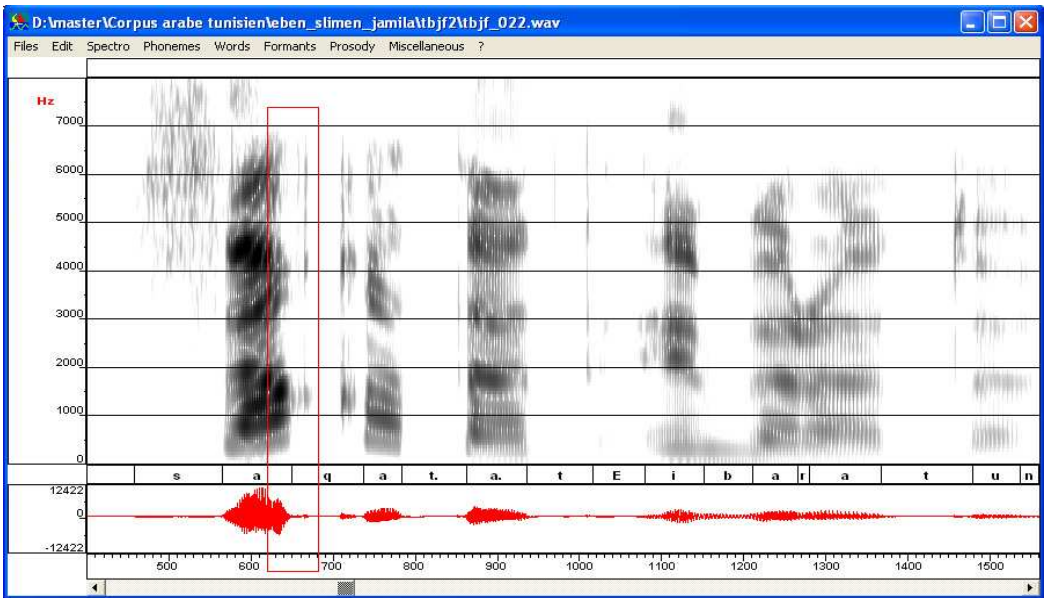


Figure 3. 6 : *Spectrogramme de l'enregistrement* «سَقَطَتْ إِبْرَةٌ.» « **Saqatat ɜibratun** » (L'aiguille est tombée) *prononcé par la locutrice Loc.W.2*

- La présence d'une voyelle emphatique qui est insérée pour compenser la durée courte du « tap » /r / et parce qu'en plus il y a une voyelle muette /°/ au dessus de la consonne /b/. Il y a donc l'insertion de cette voyelle juste après une occlusive voisée « b » pour faire percevoir le « r » (Braham.A 1997) (Voir Fig.3.7).

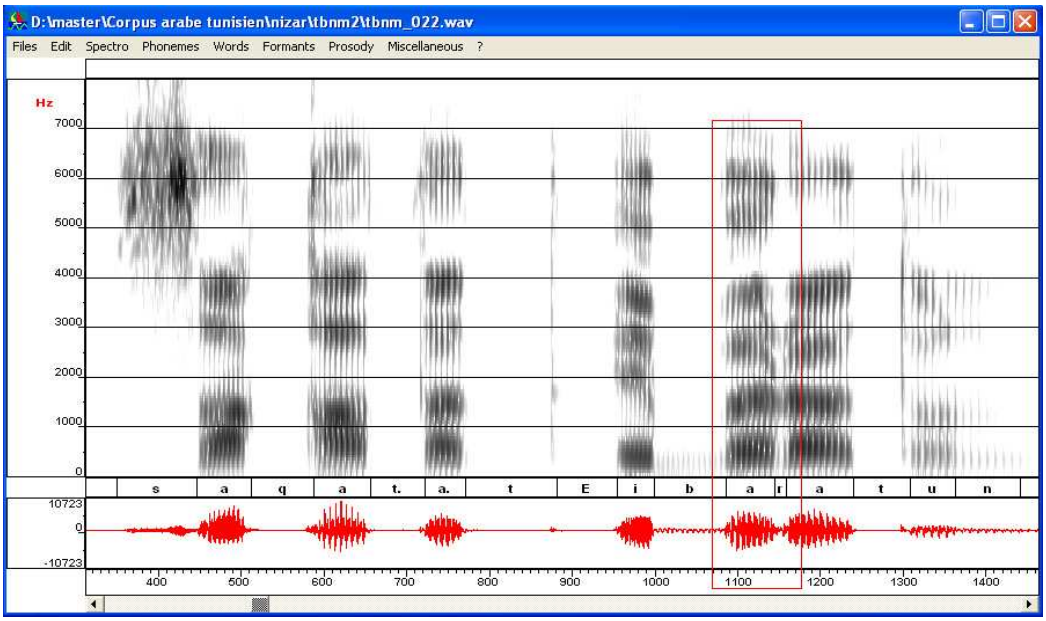


Figure 3. 7 : *Spectrogramme de l'enregistrement* «.سَقَطَتْ إِبْرَةٌ » « **Saqatat 3ibratun** » (L'aiguille est tombée) *prononcé par le locuteur Loc.M.2*

- Lorsqu' une phrase commence par une consonne occlusive non voisée, on peut prendre une durée d'occlusion par défaut de 60ms qui est la durée moyenne de l'occlusion selon le phonéticien Mr. Braham (Braham.A 1997).

En plus de ces difficultés, le locuteur d'écarte parfois de la prononciation attendue et dans ce cas nous prenons en considération ce qui a été prononcé et non pas ce qui devait être lu (taux d'erreur de 0.05%).

3.6. Etiquetage formantique

Comme il a été mentionné précédemment, nous avons effectué l'étiquetage formantique en utilisant le logiciel Winsnoori. Pour tracer les courbes des formants d'un enregistrement donné il faut tout d'abord calculer les racines LPC du signal traité à l'aide de la fonction « display LPC roots », ensuite, appliquer la fonction « keep decoration » pour coller les

racines LPC à l'image du spectrogramme. Puis, ouvrir l'interface graphique « klatt sythesizer » et suivre manuellement les points roses qui sont les racines de LPC avec le curseur de la souris ou recaler des trajectoires grossières sur ces valeurs. On obtient ainsi, les trajectoires de suivi de formants. Pour vérifier la pertinence du suivi on a eu recours à la synthèse du signal en utilisant le synthétiseur de Klatt (Rekhis.O 2009). (Voir Fig.3.8. ci-dessous)

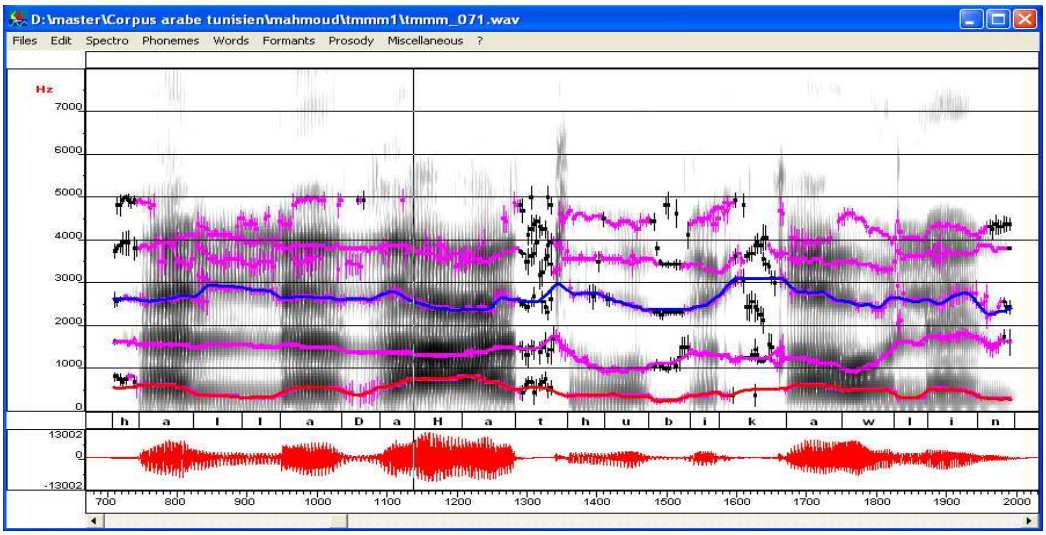


Figure 3. 8 : *Spectrogramme de l'enregistrement* «هَلْ لَذَعَتْهُ بِقَوْلٍ؟ » « **Hal laðaʕathu biqawlin** » (Est ce qu'elle l'a touché avec ses paroles ?) *prononcé par le locuteur Loc.M.1*

Au cours de cet étiquetage, nous avons rencontré plusieurs difficultés qui se concentrent en particulier sur certaines voyelles là où les formants se rapprochent fortement, pour les semi-voyelles dont les formants ne sont souvent pas visibles directement et pour certaines consonnes qui possèdent des pseudo-formants dans la continuité des formants de la voyelle précédente et ceux de la voyelle suivante au sein de la syllabe VCV (voir les exemples ci-dessous). Pour tenir compte de ces difficultés et aussi de l'absence des racines LPC dû au manque d'énergie au niveau du spectrogramme, nous avons fait appel à des experts en la matière, utilisé les valeurs nominales du phonème correspondant et écouté le signal synthétisé en utilisant le synthétiseur à formants de Klatt intégré à Winsnoori. Les tableaux 3.6, respectivement 3.7 et 3.8, représentent les valeurs nominales des formants pour les voyelles, respectivement les consonnes, retrouvées dans la littérature pour des locuteurs masculins. (Ghazali.S 1977) (Braham.A 1997).

Voyelles / Formants	F1	F2	F3
Fatha « a »	700	1500	2400
Fatha longue « A »	800	1200	2200
Dhamma « u »	300	800	2200
Dhamma longue « U »	300	2200	2700
Kasra « i »	300	2300	3200
Kasra longue « I »	300	2000	2700

Tableau 3. 6 : *Valeurs nominales des voyelles*

Pour les consonnes, on peut aussi se référer aux valeurs de (Ghazali.S 1977) (Braham.A 1997) (Voir Tableaux 3.7 et 3.8).

	F1	F2	F3
/H/			
/a/	900	1450	2300
/i/	700	1700	2700
/u/	650	1300	1700
/X/			
/a/	1100	1700	2300
/i/	700	1800	2700
/u/	550	1100	1700
/G/			
/a/	500/600	1200/1300	2300/2600
/i/			
/u/			

Tableau 3. 7 : *Valeurs nominales les consonnes /H/, /X/ et /G/* au voisinage des voyelles courtes

	/s/	**/s./**	**/t/**	**/t./**	**/D/**	**/z./**
/a/	1600	1200	1800	1300	1600	1000
/i/	2100	1500	2100	1300	1600	1100
/u/	1300	900	1100	900	1400	800

Tableau 3. 8 : *Valeurs nominales de F2 pour les consonnes /s/, /s./, /t/, /t./, /D/ et /z./* au voisinage des voyelles courtes

Les deux exemples ci-dessous illustrent certaines difficultés rencontrées au cours de l'étiquetage et les solutions trouvées pour assurer la continuité et la pertinence du suivi.

Exemple 1 :

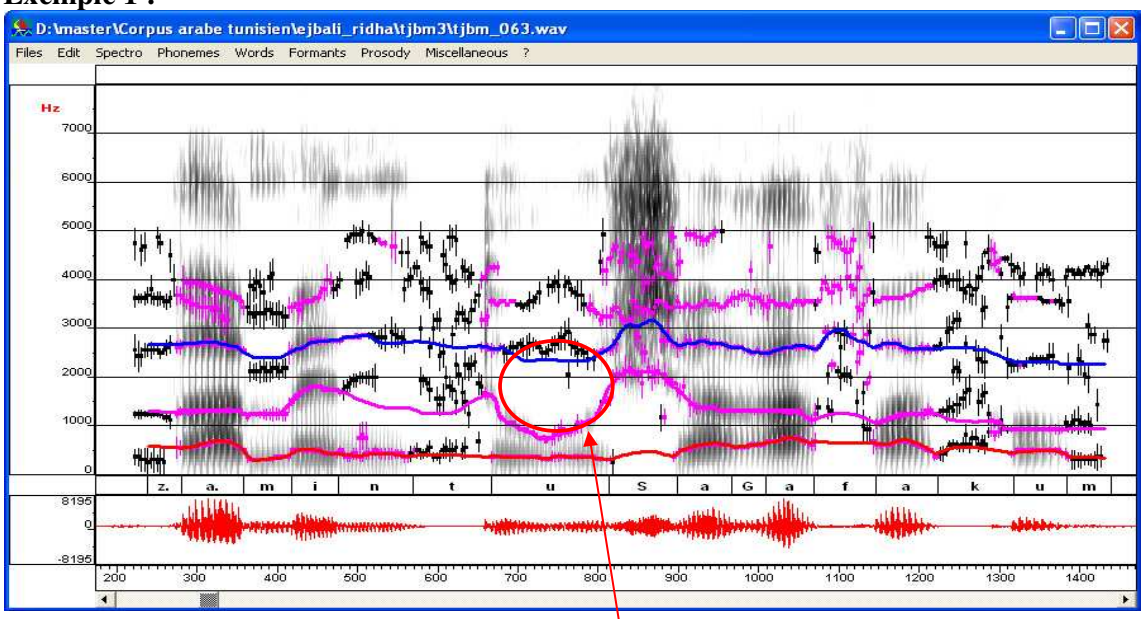


Figure 3. 9 : *Spectrogramme de l'enregistrement* «ضَمِنْتُ شَغَفَكُمْ » «**dˁamintu Šaɣafakum**» (J'ai garanti votre passion) *prononcé par le locuteur Loc.M.3*

On essaye dans les parties à faible énergie de se rapprocher des valeurs nominales.

Exemple 2 :

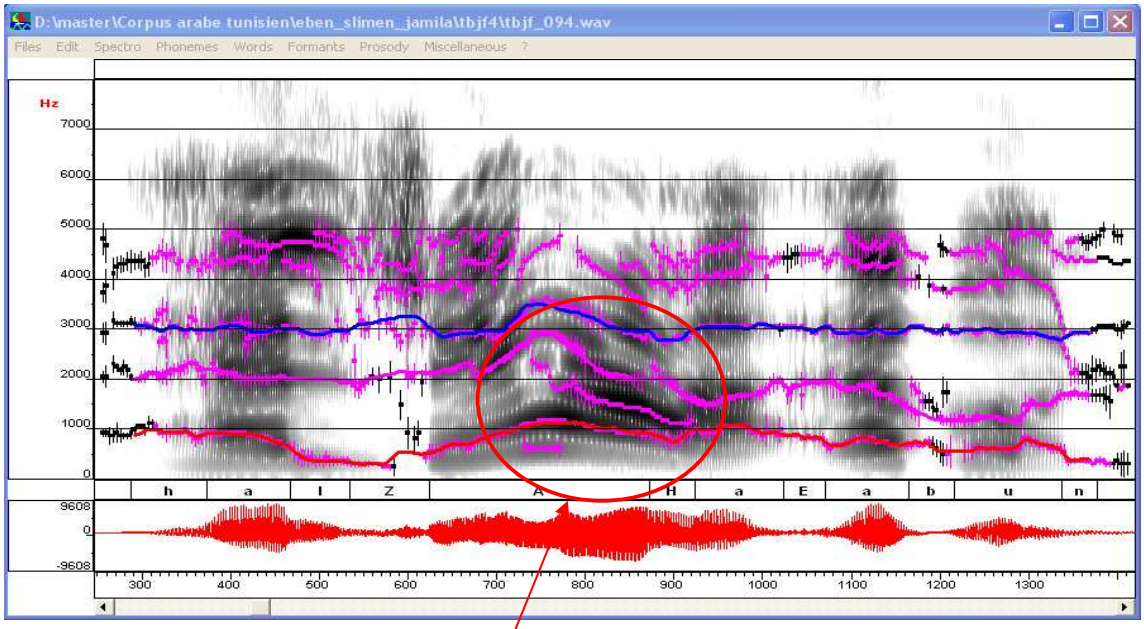


Figure 3. 10 : *Spectrogramme de l'enregistrement* « هَلْ جَاعَ أَبٌ؟ » « **Hal ja:ɛa ɜabun ?** » (Le père a-t-il faim ?) *prononcé par le locuteur Loc.F.2*

Pour lever les ambiguïtés d'étiquetage dans les parties denses en énergie, on utilise les valeurs nominales

Comme nous venons de le voir, nous avons fait appel à des experts en langue arabe pour faire les choix d'étiquetage formantique pertinent.

3.7. Conclusion

Dans ce chapitre, nous avons présenté notre corpus Arabe étiqueté manuellement qui servira de base de référence ainsi que quelques notions de linguistique et de phonétique de la langue arabe standard. Ensuite, nous avons décrit brièvement le logiciel Winsnoori avec lequel nous avons fait l'étiquetage. Enfin, nous avons terminé par une description détaillée de l'étiquetage manuel phonétique et formantique, des difficultés rencontrées au cours de cet étiquetage et des solutions prises pour y remédier et assurer la pertinence du suivi des formants de référence.

Chapitre IV
Méthodologies d'estimation et de suivi des Formants

4.1. Introduction

Le suivi de formants est destiné à retrouver les trajectoires des fréquences formantiques parallèles à l'axe de temps. Une trajectoire est une courbe représentée dans le domaine temps-fréquence. Les formants permettent aussi la détection des différents lieux d'articulation caractérisant l'évolution de la forme du conduit vocal. La représentation formantique a été utilisée essentiellement en synthèse et en modification de la parole (Alessandro.C 1992)**.** Les trajectoires formantiques sont également des paramètres pour la reconnaissance.

De nombreux efforts ont été consacrés au développement d'algorithmes de suivi de formants qui reste délicate malgré l'emploi de diverses techniques.

Dans ce chapitre, nous décrivons notre contribution qui est la méthode de suivi de formants basée sur la détection des crêtes de Fourier qui sont les maxima du spectrogramme, et autre nouvelle méthode basée sur la détection des crêtes d'ondelettes , qui sont les maxima de scalogramme en testant trois types d'ondelettes complexes, tout en utilisant comme contrainte de suivi pour les deux méthodes le calcul du centre de gravité de la combinaison des fréquences formantiques candidates. Cette contrainte assure la continuité et la robustesse du suivi. Ensuite, nous présentons notre nouvelle méthode pour l'appariement des courbes fréquentielles en utilisant la programmation dynamique combinée avec le filtrage de Kalman pour le lissage de suivi.

4.2. Approche de suivi des formants basée sur la détection des maxima locaux en utilisant le calcul de centre de gravité

Le diagramme de la figure 4.1 décrit les principales étapes de l'algorithme proposé. Chaque étape du diagramme est décrite brièvement ci-dessous.

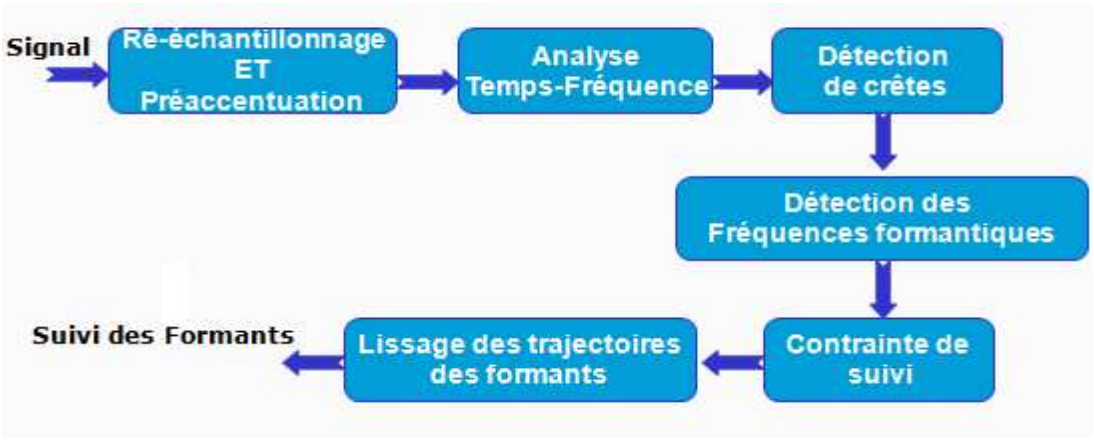


Figure 4. 1 : *Diagramme de l'algorithme de suivi des formants par détection des crêtes (maxima locaux) utilisant le calcul de centre de gravité comme contrainte de suivi*

4.2.1. Ré-échantillonnage et Préaccentuation

La fréquence d'échantillonnage utilisée dans notre base de données est 16 kHz. Puisque nous nous sommes intéressés aux trois premiers formants, le signal est ré-échantillonné à 8 kHz afin de ne pas prendre en compte des formants candidats au-dessus de 4 kHz, et nous permettre d'utiliser une analyse d'ordre inférieur plus rapide. Ensuite, le signal de parole est préaccentué par un filtre de premier ordre pour accentuer les hautes fréquences.

4.2.2. Analyse temps-fréquence

Nous présentons dans ce chapitre un nouvel algorithme de suivi de formants, nous avons besoin alors de localiser les fréquences des formants au cours de temps. C'est pour cet objectif que nous avons fait des analyses temps fréquences sur le signal de parole qui va nous permettre d'analyser la non stationnarité de ce signal.

Dans cet algorithme, nous avons appliqué séparément deux analyses temps fréquence au signal d'entrée. La première analyse est donnée par le carré du module de la transformée de Fourier fenêtrée, encore appelée transformée de Fourier à court terme (TFCT), pour obtenir le spectrogramme du signal. Le spectrogramme utilisé ici est un spectrogramme à large bande

qui lisse l'enveloppe spectrale du signal, et permet, par conséquent, de visualiser l'évolution temporelle des formants en utilisant la fenêtre d'analyse de Hamming de taille 4ms et avec un recouvrement entre les fenêtres d'analyse de 75%. Tandis que la deuxième analyse temps-fréquence est représentée par le scalogramme en appliquant la transformée en ondelettes.

On peut se demander pourquoi, si la transformée de Fourier fenêtrée fournit la représentation temps-fréquence du signal, on aurait besoin de la transformée en ondelettes.
Le problème de la transformée de Fourier fenêtrée, réside dans le principe d'incertitude d'Heisenberg. Il est impossible d'obtenir simultanément les informations de date et de fréquence d'un signal. On ne peut pas donc obtenir une représentation temps-fréquence exacte d'un signal et on ne peut pas savoir quelles composantes spectrales existent à un instant donné. Tout ce que nous pouvons connaître ce sont les intervalles de temps pendant lesquels une certaine bande de fréquence existe. Le problème de cette transformée est un problème de résolution. La difficulté est donc de choisir une fonction de fenêtrage une fois pour toutes et de l'utiliser dans toute l'analyse ; ce qui veut dire que la résolution de la transformée de Fourier fenêtrée est constante sur tout le plan temps-fréquence. Cette transformée fournit alors une analyse des signaux monorésolution. Par contre la transformée en ondelettes qui est une analyse multirésolution résout, dans une certaine mesure, le problème de la résolution comme nous allons le voir plus loin. La résolution temps fréquence obtenue en utilisant la transformée en ondelette est représentée par le scalogramme. Cette transformée en ondelette utilise l'ondelette complexe qui sépare les informations de l'amplitude et de la phase (Mallat.S, 2000).

Pour calculer le scalogramme, nous avons testé trois types d'ondelettes complexes incluses dans le toolbox de Matlab qui sont : Morlet Complexe (cmorwav), Frequency B-Spline (fbspwav) et Shanon (shanwav).

- La fonction de l'ondelette Morlet Complexe, notée par cmor, est définie par :

$$\psi(x) = \frac{1}{\sqrt{\pi f_b}} e^{2i\pi f_c x} e^{-\frac{x^2}{f_b}} \qquad \text{(Eq.4.1)}$$

où f_b est le paramètre de la largeur de bande et f_c est le centre fréquentiel de l'ondelette

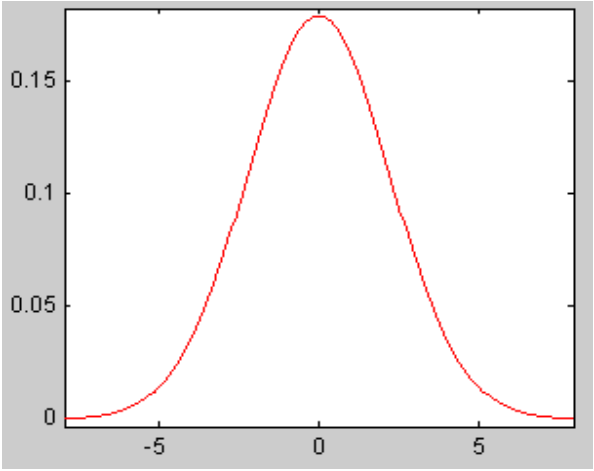


Figure 4. 2 : *Le module de la fonction Psi de l'ondelette Morlet Complexe : cmor10-1 (avec f_b=10 et f_c=1)*

- La fonction de l'ondelette Frequency B-Spline, notée par fbsp, est définie par :

$$\psi(x) = \sqrt{f_b}\left(sinc\left(\frac{f_b x}{m}\right)\right)^m e^{2i\pi f_c x} \qquad \text{(Eq.4.2)}$$

où m est un paramètre d'ordre entier de dérivation tel que $(m \geq 1)$ et f_b est le paramètre de la largeur de bande et f_c est le centre fréquentiel de l'ondelette

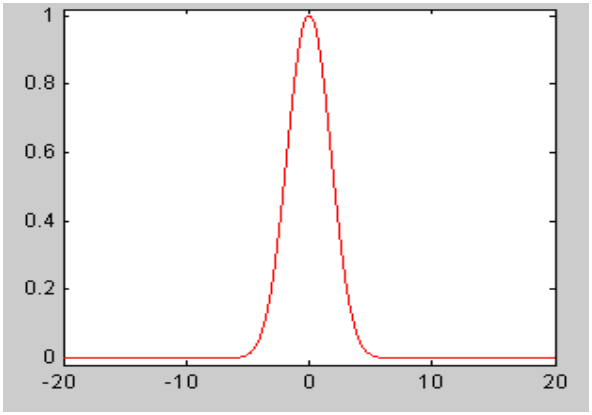


Figure 4. 3 : *Le module de la fonction Psi de l'ondelette Frequency B-Spline : fbsp10-1-1 (avec m=10, f_b=1 et f_c=1)*

- La fonction de l'ondelette Shanon, notée par shan, est définie par :

$$\psi(x) = \sqrt{f_b}\ sinc(f_b x) e^{2i\pi f_c x} \qquad \text{(Eq.4.3)}$$

où f_b est le paramètre de la largeur de bande et f_c est le centre fréquentiel de l'ondelette.

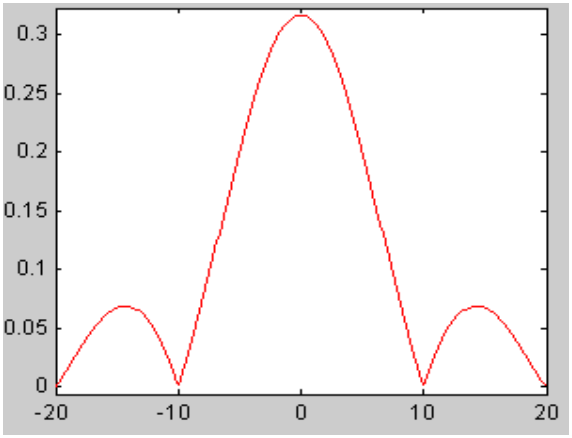


Figure 4. 4 : *Le module de la fonction Psi de l'ondelette Shanon : shan0.1-1 (avec f_b=0.1 et f_c=1)*

4.2.3. Détection de crêtes

En écoutant la musique ou bien la parole, on perçoit plusieurs fréquences qui varient dans le temps. La variation temporelle de plusieurs fréquences instantanées peut être mesurée avec des décompositions temps-fréquence et en particulier avec une transformée de Fourier à Court Terme ou transformée en ondelettes (Mallat.S 2000) (Chaplais.F 2001). L'idée de suivi est de trouver les fréquences des formants en utilisant les fréquences instantanées. En effet, on peut montrer que les maxima spectraux correspondent à des points pour lesquels la phase est stationnaire. Cela signifie qu'au voisinage d'un pic spectral, les fréquences instantanées des coefficients de la transformée de Fourier sont égales à celle du pic. L'accumulation de fréquences instantanées proches permet donc de trouver les formants. Cette propriété a notamment été utilisée par Charpentier, en appliquant une fenêtre de signal suffisamment longue, pour trouver les harmoniques de la fréquence fondamentale (Charpentier, 1986). Ici nous l'utilisons pour trouver les maxima spectraux correspondant aux formants.

Ainsi dans cette section, on va définir le signal analytique et ses propriétés, les fréquences instantanées et comment les détecter par crêtes de Fourier ou d'ondelette.

4.2.3.1. Définition des fréquences instantanées

On considère un signal *f* d'un cosinus modulé :

$$f(t) = a(t)\cos\phi(t) \quad \text{où} \quad a(t) \geq 0 \qquad \text{(Eq.4.4)}$$

La fréquence instantanée est définie comme la dérivée de la phase :

$$\omega(t) = \phi'(t) \geq 0 \qquad \text{(Eq.4.5)}$$

Le signe de la dérivée peut changer selon la variation de $\phi(t)$ en fonction du temps c'est-à-dire elle peut être soit positive, soit négative. Cependant il faut être prudent, car il existe de nombreux choix possibles pour $a(t)$ et $\phi(t)$ ce qui implique que $\omega(t)$ n'est pas définie d'une façon unique par rapport à f .

Dans ce cas, on peut utiliser la partie analytique f_a d'un signal réel f définie par :

$$f_a(t) = a(t)\exp(i\phi(t)) \qquad \text{(Eq.4.6)}$$

f_a peut se décomposer proprement en module et phase complexe.

où $a(t)$ est l'amplitude analytique de f et $\phi'(t)$ est sa fréquence instantanée comme définie précédemment dans l'équation (4.5).

La principale condition pour qu'un signal soit analytique est que sa transformée de Fourier soit nulle sur les fréquences négatives.

Les signaux analytiques sont des outils précieux de synthèse de signaux possédant un contenu fréquentiel instationnaire précis.

4.2.3.2. Détection des crêtes de Fourier

L'algorithme de crêtes calcule les fréquences instantanées à partir des maxima locaux du spectrogramme. Puisque les fréquences des formants varient lentement au cours de temps, il est possible de les considérer constant durant une fenêtre d'analyse auquel on applique la transformée de Fourier.

En effet, la famille des atomes de la transformée de Fourier fenêtrée notée $g_{u,\xi}$ est générée par des translations temporelles et des modulations fréquentielles d'une fenêtre réelle et symétrique $g(t)$ de type Hamming et de taille 4 ms avec un chevauchement de 75%. Cet atome a comme fréquence centrale ξ et il est symétrique par rapport à u facteur de translation. Le spectrogramme $Psf(u,\xi) = |Sf(u,\xi)|^2$ mesure l'énergie de signal f au voisinage de l'atome (u,ξ) dans le domaine temps fréquence et où $Sf(u,\xi)$ est la transformée de Fourier fenêtrée par la corrélation de cet atome avec le signal d'entée *f*. L'algorithme

calcule les fréquences instantanées de ce signal qui sont considérées comme maxima locaux de son spectrogramme.

Considérons un signal analytique *f(t)* tel que *f(t)= a(t) cos ϕ(t)* avec *a(t)* l'amplitude analytique et *ϕ'(t)* est sa fréquence instantanée. *a(t)* et $\phi'(t)$ varient lentement au cours de temps. Il a été démontré de point de vue théorique dans (Mallat.S, 2000) que la fréquence instantanée de *f* est reliée à la transformée de Fourier fenêtrée $Sf(u,\xi)$ si $\xi \geq 0$ par la relation suivante (Eq.4.7.) :

$$Sf(u,\xi) = \frac{\sqrt{s}}{2} a(u) \exp\, i(\phi(u) - \xi(u)) \times \left[\hat{g}\left(s\left[\xi - \phi'(u)\right]\right) + \varepsilon(u,\xi) \right] \qquad \text{(Eq.4.7)}$$

où *s* est une échelle appliquée sur la fenêtre de Fourier *g*, $\hat{g}$ est la transformée de Fourier de g et $\varepsilon(u,\zeta)$ est le terme correctif. Comme $|\hat{g}(\omega)|$ est maximum en $\omega = 0$, l'équation (Eq.4.7) montre que pour chaque *u,* le spectrogramme est maximum en sa fréquence centrale $\xi(u) = \phi'(u)$. Donc on peut conclure que les fréquences instantanées sont validées en tant que maxima locaux de spectrogramme. Les maxima aux points $(u, \xi(u))$ forment donc les crêtes de Fourier du plan temps-fréquence. A chaque instant de la fenêtre d'analyse, l'algorithme détecte tous les maxima locaux de la représentation temps fréquence assimilés aux crêtes de Fourier dans le plan des points $(u, \xi(u))$ et c'est ainsi qu'on obtient pour chaque formant la combinaison des fréquences candidates.

4.2.3.3. Détection des crêtes d'ondelettes

Les atomes de Fourier fenêtrée ont une taille fixe et ne peuvent donc pas suivre fidèlement des fréquences instantanées variant à des vitesses très différentes. Par contre, la transformée en ondelettes analytiques modifie l'échelle de ses atomes temps-fréquences. L'algorithme de crêtes est étendu à la transformée en ondelette analytique afin de mesurer avec exactitude des fréquences instantanées variant plus rapidement en haute fréquence.

Pour cela considérons dans notre étude une ondelette complexe définie par (Mallat.S, 2000) (Chaplais.F, 2001) :

$$\psi(t) = g(t)\exp(i\eta t) \qquad \text{(Eq.4.8)}$$

Avec η le centre de fréquence de l'ondelette.

Comme pour la transformée de Fourier fenêtrée, g est une fenêtre symétrique de support [-1/2,1/2] et normalisée à l'ordre de 1.

Soit $\Delta\omega$ la largeur de bande de $\hat{g}$ définie par :

$$\left|\hat{g}(\omega)\right| << 1 pour |\omega| \geq \Delta\omega \qquad \text{(Eq.4.9)}$$

$$\text{Si } \eta > \Delta\omega \text{ alors } \forall \omega < 0, \hat{\psi}(\omega) = \hat{g}(\omega - \eta) << 1 \qquad \text{(Eq.4.10)}$$

La famille des atomes temps fréquence notée $\psi_{u,s}$ est obtenue de l'atome de base ψ par dilatation selon l'échelle *s* et par la translation selon le paramètre *u*. La fonction $\psi_{u,s}$ est centrée en *u* comme l'atome de Fourier fenêtrée mais l'échelle *s* est un paramètre variable qui décrit $\Re^+$ et $\xi = \eta / s$ est le centre de fréquence de l'ondelette dilatée.

De même, il a été démontré qu'il y a une relation entre la fréquence instantanée de $f(t) = a(t)\cos(\phi(t))$ et $Wf(u,s)$ la transformée en ondelette, définie par :

$$Wf(u,s) = \frac{\sqrt{s}}{2} a(u) \exp[i\phi(u)] \left[\hat{g}(s(\xi - \phi'(u)) + \varepsilon(u,\xi) \right] \qquad \text{(Eq.4.11)}$$

Le terme correctif $\varepsilon(u,\xi)$ est négligeable si a (t) et $\phi'(t)$ varient lentement sur le support de $\psi_{u,s}$ et si $\phi'(u) \geq \Delta\omega / s$.

La fréquence instantanée peut être mesurée à partir des crêtes de la transformée en ondelettes en utilisant un scalogramme normalisée défini par :

$$\frac{\xi}{\eta} P_w f(u,\xi) = \frac{|Wf(u,s)|^2}{s} \text{ pour } \xi = \eta / s \qquad \text{(Eq.4.12)}$$

Pour le signal $f(t)$ le scalogramme normalisée se calcule à l'aide de :

$$\frac{\xi}{\eta} P_w f(u,\xi) = \frac{1}{4} a^2(u) \left| \hat{g}\left(\eta \left[1 - \frac{\phi'(u)}{\xi} \right] \right) + \varepsilon(u,\xi) \right|^2 \qquad \text{(Eq.4.13)}$$

Comme $\left|\hat{g}(\omega)\right|$ est maximum en $\omega = 0$, cette expression montre que si $\varepsilon(u,\xi)$ est négligeable alors le scalogramme atteint son maximum lorsque $\frac{\eta}{s(u)} = \xi(u) = \phi'(u)$.

Les maxima $(u,\xi(u))$ au scalogramme normalisé forment les crêtes d'ondelettes dans le plan temps-fréquence. Sous les mêmes conditions que le spectrogramme, elles indiquent les fréquences instantanées dans la limite de la résolution de la transformée.

L'amplitude analytique est donnée par :

$$a(u)=\frac{2\sqrt{\eta^{-1}\xi PWf(u,\xi)}}{\left|\hat{g}(0)\right|} \qquad \text{(Eq.4.14)}$$

La phase complexe de $Wf(u,s)$ vaut $\phi_w(u,\xi)=\phi(u)$ aux points de crête :

$$\frac{\partial\phi_w(u,\xi)}{\partial(u)}=\phi'(u)=\xi \qquad \text{(Eq.4.15)}$$

Dans le cas où f est la somme de deux composantes spectrales, il faut vérifier que les fréquences instantanées ne créent pas des interférences qui détruisent la structure des crêtes pour cela, il faut que :

$$\frac{|\phi'_1(u)-\phi'_2(u)|}{\phi'_1(u)}\geq\frac{\Delta\omega}{\eta} \text{ et } \frac{|\phi'_1(u)-\phi'_2(u)|}{\phi'_2(u)}\geq\frac{\Delta\omega}{\eta} \qquad \text{(Eq.4.16)}$$

où $\Delta\omega$ est la largeur de bande de $\hat{g}$ et η est le centre fréquentiel de l'ondelette. Ceci montre que pour séparer des fréquences instantanées voisines, la largeur de bande relative $\frac{\Delta\omega}{\eta}$ doit être suffisamment petite. La largeur de bande $\Delta\omega$ de $\hat{g}$ est une constante de l'ordre de1, mais η est un paramètre que l'on choisit afin de satisfaire à la fois les conditions temporelles et les conditions fréquentielles.

Ces conditions portent sur les écarts relatifs en fréquence. Elles sont liées au type du pavage du plan temps-fréquence.

Dans ce cas, on peut conclure que les crêtes d'ondelette constituent un outil de détection de fréquences instantanées lorsque celles-ci ne sont pas trop proches et l'algorithme détecte toutes les fréquences instantanées du signal d'entrée f qui sont considérées comme maxima locaux de son scalogramme. A chaque instant de la fenêtre d'analyse, l'algorithme détecte tous les maxima locaux de la représentation temps fréquence aux points $(u,\xi(u))$ assimilés aux crêtes d'ondelette et c'est ainsi qu'on obtient pour chaque formant la combinaison des fréquences candidates.

4.2.4. Détection des fréquences formantiques

On détecte donc tous les maxima locaux du spectrogramme, respectivement du scalogramme. Puisque on ne considère que seulement les trois premiers formants, on récupère les points de crête à chaque instant et on procède à un seuillage selon trois largeurs de bande pour chaque formant pour mieux localiser les crêtes de Fourier, respectivement d'ondelette. Ainsi on obtient pour chaque formant la combinaison de fréquences candidates tout en éliminant les valeurs aberrantes, les crêtes de faibles amplitudes. Ces valeurs peuvent être des artefacts provenant par exemple, des « ombres » d'autres fréquences produites par les lobes latéraux de la fenêtre de transformée de Fourier, respectivement de la fenêtre de l'ondelette ou encore des fréquences instantanées spécifiques à la fréquence fondamentale F0 (Châari.S, 2006).

4.2.5. Contrainte de suivi

Généralement, on considère que les formants varient lentement au cours de temps ce qui nous mène à imposer une contrainte de suivi dans le processus de sélection des fréquences de chaque formant. Dans cette étude, nous proposons de calculer le centre de gravité de la combinaison candidate correspondante à chaque formant. A l'intérieur de chaque largeur de bande et à chaque instant, nous calculons la valeur de chaque formant en prenant le centre de gravité de l'ensemble des fréquences instantanées correspondante pondérées par l'énergie spectrale :

$$\bar{f} = \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i} \qquad \text{(Eq.4.17)}$$

où f_i fréquence du i^{ème} candidat, p_i son énergie spectrale et n le nombre total de valeurs de fréquences instantanées considérées.

On calcule le centre de gravité de l'ensemble des fréquences formantiques candidates détectées par l'étape précédente de l'algorithme et la valeur trouvée sera la fréquence formantique recherchée. On passe ensuite à la combinaison candidate suivante, ce qui permet de récupérer les points de suivi point par point pour chaque formant.

4.2.6 Lissage des trajectoires des formants

Pour bien lisser les trajectoires des formants, on interpole la suite des points résultants de la fonction (Eq4.17) précédente dans l'étape de calcul de la contrainte de suivi. Ensuite pour lisser les trajectoires des formants, on a calculé la valeur moyenne de déplacement de chaque point de la trajectoire correspondant à un formant, avec les fréquences déjà choisies précédemment c'est-à-dire en gardant à chaque fois l'historique de suivi. Suite à cette étape, les trajectoires correspondant aux trois premiers formants sont continues et lisses.

4.3. Approche de suivi de formants basée sur la programmation dynamique combinée avec le filtrage de Kalman

Le diagramme de la figure 4.5 décrit les principales étapes de l'algorithme proposé. Chaque étape du diagramme est décrite brièvement ci-dessous.

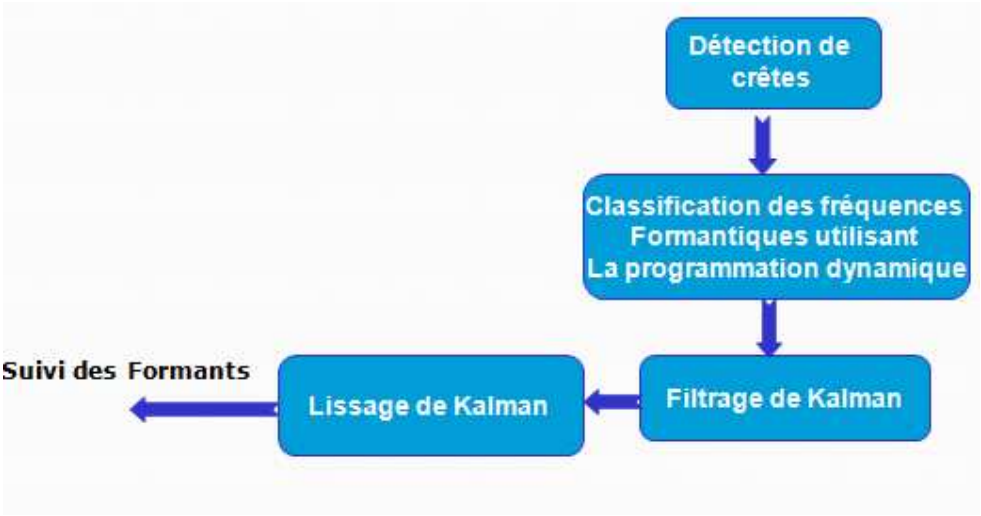


Figure 4. 5 : *Diagramme de l'algorithme de suivi des formants par détection des crêtes basé sur la programmation dynamique en combinaison avec le filtrage de Kalman*

4.3.1. Détection de crêtes

Comme on l'a déjà expliqué précédemment, le signal d'entrée doit être d'abord ré-échantillonner à 8 kHz. Ensuite, nous avons préaccentué le signal de parole par un filtre de premier ordre pour accentuer les hautes fréquences. Puis, pour détecter les crêtes nous avons appliqué l'analyse temps fréquence sur le signal; il s'agit de calculer le spectrogramme par

transformée de Fourier fenêtrée si on va détecter les crêtes de Fourier et calculer le scalogramme par transformée d'ondelette complexe si on va détecter les crêtes d'ondelette. Suite à ces étapes, nous arrivons au stage de la détection des crêtes pour fournir l'ensemble des fréquences candidates de formants. Les différentes opérations de cette étape sont bien expliquées dans la section de l'approche précédente. Une fois nous avons détecté les crêtes, c'est-à-dire détecter à chaque instant tous les maxima locaux de la représentation temps fréquence du signal d'entrée, nous fournissons le vecteur temps et le vecteur de fréquence correspondant dans deux fichiers texte comme données d'entrée pour l'étape suivante de l'algorithme.

4.3.2. Classification des fréquences formantiques utilisant la programmation dynamique

Suite à l'étape de la détection des fréquences candidates, nous avons besoin de les classifier selon chaque formant. Pour cela nous nous sommes basés sur le principe de la programmation dynamique et de calcul de coût de probabilité de transition en intégrant des contraintes de continuité pour avoir une meilleure classification.

En implémentant cet algorithme, nous nous sommes inspirés de celui de Depalle dans (Depalle.Ph, 1992) (Deppalle.Ph, 1993). Tout d'abord l'algorithme commence par lire les fichiers texte des données d'entrée contenant le vecteur temps et le vecteur des fréquences candidates et les mettre dans une table de type double. Ensuite, nous avons implémenté une fonction dans cet algorithme qui parcourt la table de données à travers une fenêtre de taille 3. Il s'agit de prendre à chaque analyse une fenêtre de trois vecteurs (trames) de fréquences à trois instants successives qui sont : (k-2), (k-1) et (k). Cette fonction parcourt alors toutes les données et calcule le nombre total des fenêtres. Ensuite, à chaque fenêtre elle parcourt les trois trames pour savoir le nombre maximum de formants sur la fenêtre actuelle par exemple si on a les trois trames suivantes on aura le nombre maximal de formant égal à 3 :

$$\begin{bmatrix}1000\\500\\200\end{bmatrix}\begin{bmatrix}500\\200\\100\end{bmatrix}\begin{bmatrix}500\\100\end{bmatrix}$$

(k-2) (k-1) (k)

Par la suite, à la fenêtre correspondante nous calculons toutes les combinaisons possibles de chaque trame en donnant à chaque pic son indice. Par exemple si on a une trame qui est formée de trois pics toutes les combinaisons possibles d'indices seront :

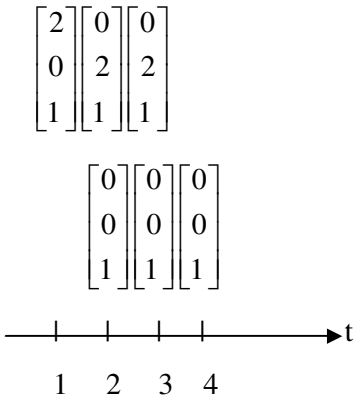
$$\begin{bmatrix}0\\0\\1\end{bmatrix}\begin{bmatrix}0\\1\\0\end{bmatrix}\begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}0\\2\\1\end{bmatrix}\begin{bmatrix}2\\0\\1\end{bmatrix}\begin{bmatrix}2\\1\\0\end{bmatrix}\begin{bmatrix}3\\2\\1\end{bmatrix}$$

Ici l'indice 1 présente le premier formant, l'indice 2 présente le deuxième formant, l'indice 3 présente le troisième formant et l'indice 0 présente un pic parasite.

Cette opération se fait bien sûr pour toutes les trames à cette fenêtre actuelle et en considérant deux états notés S_{k-1} et S_k. Chaque état est formé de deux trames. Il s'agit de calculer toutes les combinaisons possibles des deux états S_{k-1} et S_k.

Les problèmes de complexité de calculs et de volume de mémoire dus au grand nombre d'états possibles, nous ont forcés à introduire des contraintes sur les combinaisons des indices dans chaque trame. Ces contraintes de continuité sont définies comme suit :

- Interdire les naissances et les morts de trajets dans les états c'est-à-dire au sein d'une même fenêtre d'analyse de taille 3 contenant trois trames de pics à apparier, on trouve les trajets optimaux dans cette fenêtre, à l'aide de la programmation dynamique en calculant le meilleur coût. Cela implique que pour qu'un seul trajet soit détecté dans la fenêtre, il doit avoir une longueur minimale correspondant à 3 instants d'analyse et ni sa naissance ni sa mort ne doivent se trouver à l'intérieur de la fenêtre. Ensuite, en faisant glisser cette fenêtre avec un pas d'un instant d'analyse, on refait la même chose et on peut trouver de nouvelles trajets qui naissent et d'autres qui n'existent plus. Les morts des trajets sont détectées en mémorisant les trajets trouvés dans la fenêtre précédente et en cherchant ceux qui ont disparu dans la fenêtre actuelle. La mort de ces trajets s'est alors produite au dernier instant de la fenêtre précédente. Par exemple pour deux fenêtres d'analyse, on peut trouver les deux résultats successifs suivants :



Ce qui implique que le trajet d'indice 2 est mort à l'instant t=3.

- Les indices des différents pics d'une même trame doivent être ordonnés dans l'ordre croissant et on ne doit pas avoir de croisements entre les trajets.

  Par exemple les transitions possibles effectuées d'un état S_{k-1} donné vers l'état S_k est :

  L'état $S_{k-1} = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ ne peut effectuer une transition que vers les états suivants :

  $$S_k = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} ou\ S_k = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} ou\ S_k = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

Une fois nous avons généré toutes les combinaisons candidates au sein d'une même fenêtre d'analyse, l'algorithme calcule le coût de chaque chemin en utilisant la fonction de coût ci-dessous en se référant à l'équation utilisée par Depalle dans (Deppalle.Ph, 1993) :

$$\theta_k(j) = 1 - (1-\mu)\exp\left\{-\frac{\left[\Delta f_k(j,r) - \Delta f_{k-1}(r,t)\right]^2}{\sigma^2}\right\} \qquad \text{(Eq.4.18)}$$

Où *j*, *r* et *t* sont trois différents pics correspondants respectivement aux 3 trames d'instants k, (k-1) et (k-2), μ et σ sont des constantes, $\Delta f_k(j,r) = f_k(j) - f_{k-1}(r)$ et $\Delta f_{k-1}(r,t) = f_{k-1}(r) - f_{k-2}(t)$.

Une fois on a calculé tous les coûts de tous les chemins possibles dans la fenêtre d'analyse actuelle, on récupère après le chemin optimal celui qui a le meilleur coût (coût minimum) et ce chemin sera sauvegarder dans une liste. On glisse à chaque fois la fenêtre avec un pas d'instant d'analyse pour refaire les mêmes étapes jusqu'à la fin du signal traité.

Finalement, nous avons implémenté une fonction « backtracking » qui récupère la liste des chemins optimaux et construit les trajets en marche arrière pour assurer la continuité.

4.3.3. Filtrage de Kalman

L'algorithme de suivi utilise les équations de filtrage de Kalman où les mesures sont les fréquences de résonance choisies par l'étape précédente basée sur la programmation dynamique. La représentation du modèle du système dynamique filtre de Kalman est donnée comme suit :

$$x(k+1) = Fx(k) + w(k) \qquad \text{(Eq.4.19)}$$

$$y(k) = Hx(k) + v(k) \qquad \text{(Eq.4.20)}$$

où (Eq.4.19) est l'équation d'état, (Eq.4.20) est l'équation d'observation, x (k) est le vecteur d'état caché à l'instant k, y(k) est le vecteur d'observation (c'est le vecteur de mesure obtenu par les résonances candidates), F est la matrice de prédiction d'états, H est la matrice d'observations et w(k) et v(k) sont des bruits blanc gaussiens non corrélés de covariances respectives Q et R.

Selon la théorie proposée ici, nous supposons que les trajectoires des fréquences de résonance peuvent être modélisées approximativement comme des fonctions linéaires du temps dans des intervalles courts (Özbek.I.Y, 2006). Les changements non linéaires de la trajectoire (changement brusque dans les trajectoires) sont modélisés via un processus de bruit du système dynamique représenté ci-dessus. Donc les matrices *F*, *H*, *Q* et *R* sont indépendantes de paramètre temps et sont définies comme suit :

$$F=\begin{bmatrix}1&0&1&0\\0&1&0&1\\0&0&1&0\\0&0&0&1\end{bmatrix},\ H=\begin{bmatrix}1&0&0&0\\0&1&0&0\end{bmatrix},\ Q=\begin{bmatrix}0.0100&0&0&0\\0&0.0100&0&0\\0&0&0.0100&0\\0&0&0&0.0100\end{bmatrix},\ R=\begin{bmatrix}0.1000&0\\0&0.1000\end{bmatrix} \qquad \text{(Eq.4.21)}$$

L'espérance conditionnelle $\hat{x}_k = \mathrm{E}[x_k \mid y_{1:k}]$ et sa covariance $B_k = \mathrm{E}\left[\left(x_k - \hat{x}_k\right)\left(x_k - \hat{x}_k\right)^T\right]$ permettent une estimation meilleure de x_k connaissant toutes les

données jusqu'à l'instant *k*. Elles sont calculées récursivement par le filtre de Kalman et définissent entièrement la distribution de probabilité puisque $x_k \left| y_{1:k} \right.$ suit une loi gaussienne (Papadakis.N, 2007).

Selon Papadakis (Papadakis.N, 2007), connaissant l'état initial du vecteur d'état noté x_0 et l'état initial de la covariance B_0, on suppose qu'on a :

$$\hat{x}_0 = x_0 \qquad \text{(Eq.4.22)}$$

Ce qui permet d'obtenir un modèle de propagation de l'état et de sa covariance d'erreur :

$$\hat{x}_{k+1|k} = F_k \hat{x}_k \qquad \text{(Eq.4.23)}$$

$$B_{k+1|k} = Q_k + F_k B_k F_k^T \qquad \text{(Eq.4.24)}$$

Ainsi qu'une innovation de l'état et de sa covariance d'erreur :

$$\hat{x}_k = \hat{x}_{k|k-1} + K_k \left(y_k - H_k \hat{x}_{k|k-1} \right) \qquad \text{(Eq.4.25)}$$

$$B_k = B_{k|k-1} - K_k H_k B_{k|k-1} \qquad \text{(Eq.4.26)}$$

$$\text{où } K_k = B_{k|k-1} H_k^T \left(R_k + H_k B_{k|k-1} H_k^T \right)^{-1} \qquad \text{(Eq.4.27)}$$

est la matrice de gain du filtre de Kalman.

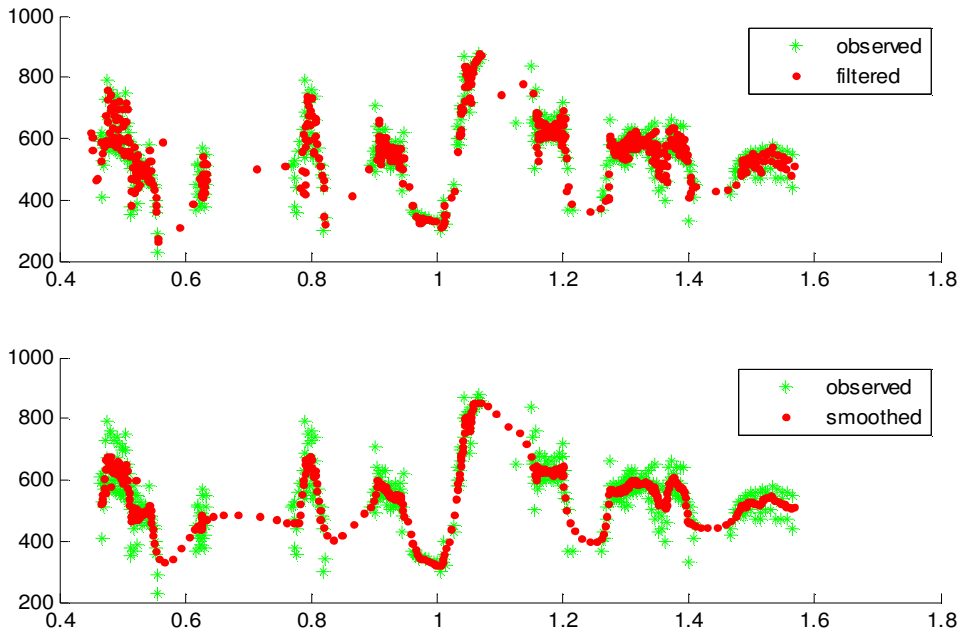
Cette dernière est multipliée par l'innovation $y_k - H_k \hat{x}_{k|k-1}$ (différence entre la mesure observée et la mesure prédite) durant la phase de correction de la trajectoire. (Papadakis.N, 2007).

4.3.4. Lissage de Kalman

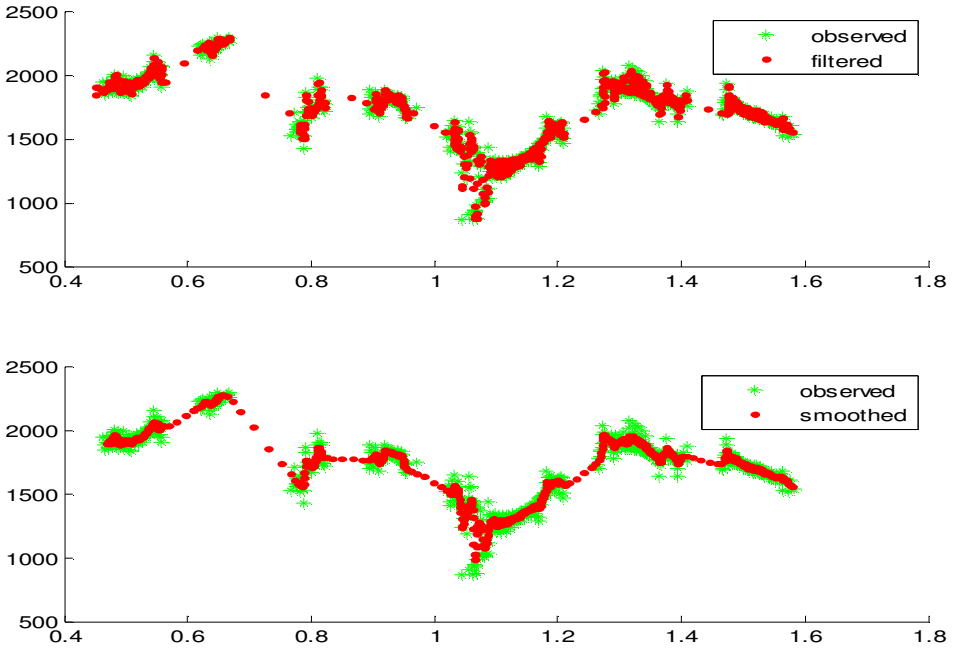
Il permet de lisser les trajectoires de suivi détecté ci-dessus. Le but est de chercher $\hat{x}_k = \mathrm{E}\left[x_k \left| y_{1:N} \right. \right]$ où *N* > *k*. Une fois l'estimation $\hat{x}_k^1 = \mathrm{E}\left[x_k \left| y_{1:k} \right. \right]$ obtenu, il ne reste donc qu'à déterminer $\hat{x}_k^2 = \mathrm{E}\left[x_k \left| y_{k+1:N} \right. \right]$ pour obtenir le lissage souhaité. Le filtrage de Kalman est donc appliqué deux fois ce qui permet de d'obtenir $\hat{x}_k^1$ par filtrage "avant" et $\hat{x}_k^2$ par filtrage

"arrière". Il s'agit de le réaliser de manière rétrograde du temps N au temps k, en usant les matrices de covariance de modèle et de mesure associées. Un autre filtre de Kalman incorporant les lois des deux estimations est par la suite utilisée, une fois elles sont combinées (Papadakis.N, 2007).

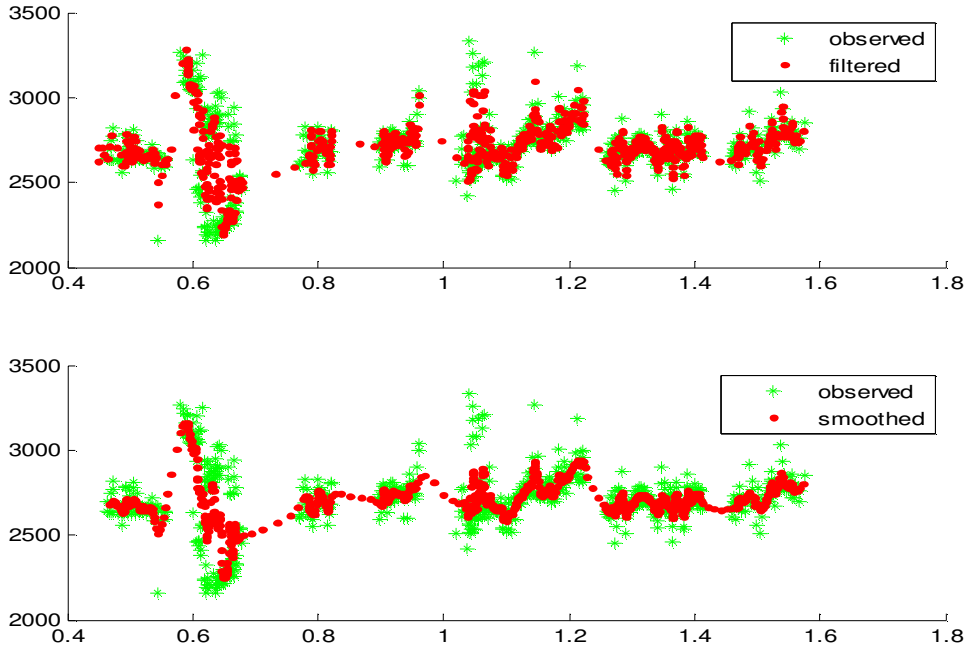
Les résultats de filtrage et de lissage de Kalman des trajectoires de chaque formant sont présentés par la figure 4.6 pour un signal de parole donné. Le résultat de suivi final correspondant de ce signal est présenté par la figure 4.7.



(a)



(b)



(c)

Figure 4. 6 : *Résultats de filtrage et de lissage de Kalman appliqués sur les trois premiers formants du signal sonore « .أَلِفَتِ العَنَانَ» «**ʒalifati lɛana:na** » prononcée par le locuteur M2 (a)filtrage et lissage de Kalman sur F1, (b) filtrage et lissage de Kalman sur F2, (c) filtrage et lissage de Kalman sur F3.*

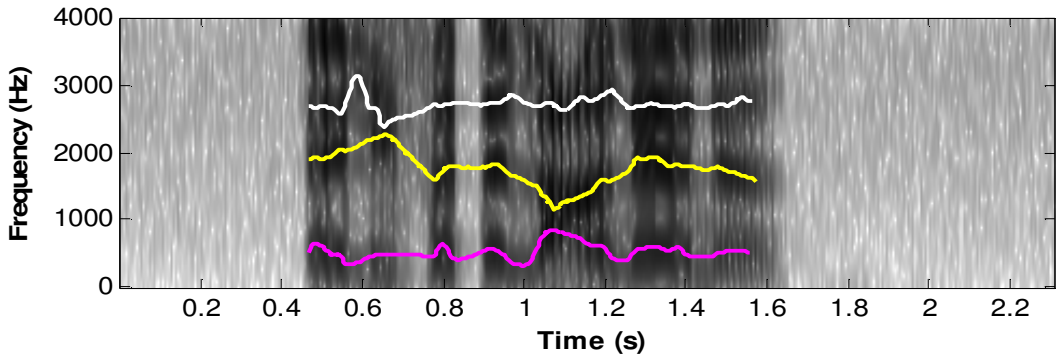


Figure 4. 7 : *Trajectoires formantiques estimées par prédiction des crêtes de Fourier utilisant le filtrage de Kalman du signal sonore « أَلِفَتِ العَنَانَ. » «**ʒalifati lɛana:na** »* (Elle s'est habituée au nuage) *prononcée par le locuteur M2*

4.4. Conclusion

Nous avons présenté dans ce chapitre notre contribution qui est la méthode de suivi de formants basée sur la détection des crêtes de Fourier, et une autre méthode nouvelle basée sur la détection des crêtes d'ondelettes qui sont les maxima de scalogramme tout en utilisant comme contrainte de suivi pour les deux méthodes le calcul du centre de gravité de la combinaison des fréquences formantiques candidates. Ensuite, nous avons présenté une autre nouvelle méthode de suivi en utilisant la programmation dynamique combinée avec le filtrage de Kalman pour le lissage de suivi.

Nous allons explorer les deux nouvelles approches de suivi dans le prochain chapitre en faisant des tests sur des signaux synthétiques et d'autres réels en prenant les signaux étiquetés manuellement issus de notre base élaborée comme référence.

Chapitre V
Application, tests et résultats

5.1. Introduction

Il est bien établi que le spectrogramme de parole fournit une information phonétique importante, en particulier les pics spectraux qui correspondent à des maxima du spectrogramme.

En plus du spectrogramme traditionnel qui sont basés sur la transformation de Fourier, d'autres représentations du signal de parole ont reçues une grande attention dans le comité de recherche ces dernières années, telle que la transformée en ondelette qui calcule la représentation temps fréquence connue par le scalogramme.

Dans ce chapitre, nous allons présenter les résultats de notre nouvelle approche basée sur les crêtes d'ondelettes en utilisant le calcul de centre de gravité sur des signaux synthétiques puis ensuite sur des signaux réels de notre corpus étiqueté. Nous allons faire une évaluation quantitative de cette approche en la comparant avec d'autres méthodes automatiques de suivi de formants et en prenant les signaux étiquetés issus de la base élaborée comme référence tout en montrant l'utilité de notre base de données étiquetées manuellement. Finalement, nous allons faire aussi une évaluation quantitative de notre deuxième nouvelle approche basée sur la programmation dynamique combinée avec le filtrage de Kalman en prenant les mêmes signaux étiquetés utilisés précédemment tout en comparant cette approche avec d'autres méthodes automatiques de suivi de formants.

5.2. Tests sur les signaux synthétiques et interprétations

Nous avons testé notre approche du suivi de formants basée sur la détection des crêtes d'ondelettes en utilisant le calcul de centre de gravité sur des voyelles synthétiques pour se rapprocher de la parole voisée naturelle et pour évaluer notre travail avant de passer aux signaux réels. L'avantage des voyelles synthétiques est le fait de connaître d'avance les formants de chaque voyelle ce qui nous aide à valider notre travail. Ces voyelles synthétiques ont été construites à l'aide d'une fonction qui s'appelle « Makevowel » qui se trouve sur le site officiel « Mathworks ». Pour cela nous avons utilisé trois voyelles, à savoir /a/, /i/, /u/ avec une fréquence d'échantillonnage de 8000 Hz et le pitch de chacune est pris égal à 100 Hz. Pour calculer le scalogramme, nous avons testé trois types d'ondelettes complexes qui sont : Morlet Complexe (cmor), Frequency B-Spline (fbsp) et Shanon (shan). Tout d'abord, nous avons fait un petit test pour les trois ondelettes afin de fixer les paramètres appropriés pour chaque ondelette, telles que les fréquences f_b et f_c, pour donner une bonne qualité de scalogramme et un bon suivi : c'est le compromis entre la résolution de l'image et l'estimation des paramètres du scalogramme.

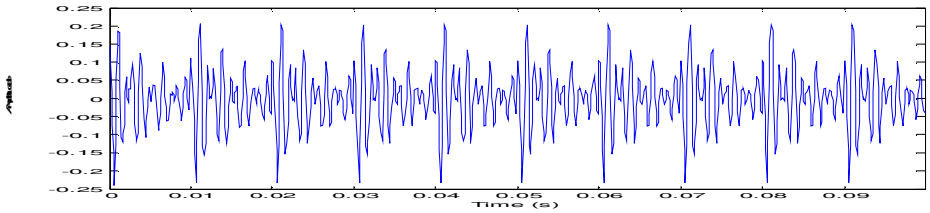
Finalement, nous avons retenu :

- pour l'ondelette cmor : f_b=10 et f_c=1
- pour l'ondelette fbsp : m=10, f_b=1 et f_c=1
- pour l'ondelette shan : f_b=0.1 et f_c=1

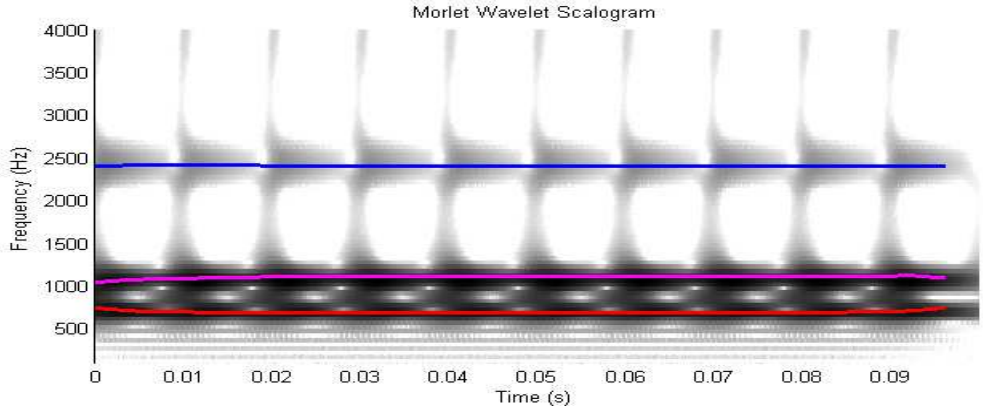
- La voyelle /a/

Les valeurs des trois premiers formants F1, F2 et F3 caractérisant la voyelle /a/ sont respectivement : 730 Hz, 1090 Hz et 2440 Hz.

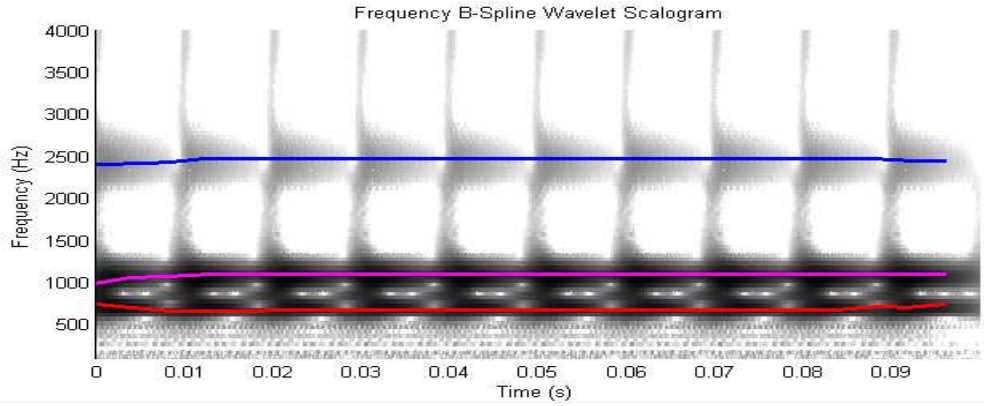
La représentation temporelle de cette voyelle, ainsi que les suivis automatiques des formants basés sur la détection des crêtes d'ondelettes dans le plan temps-fréquence en testant les trois types d'ondelettes sont illustrés par les figures suivantes.



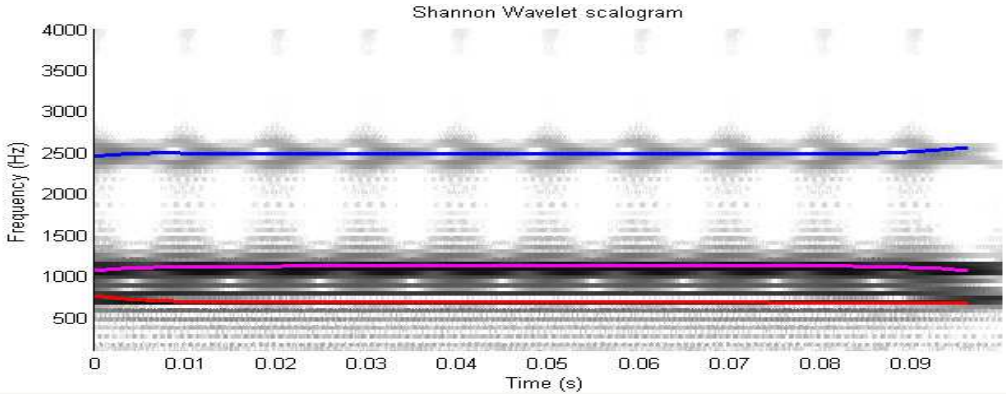
(a)



(b)



(c)



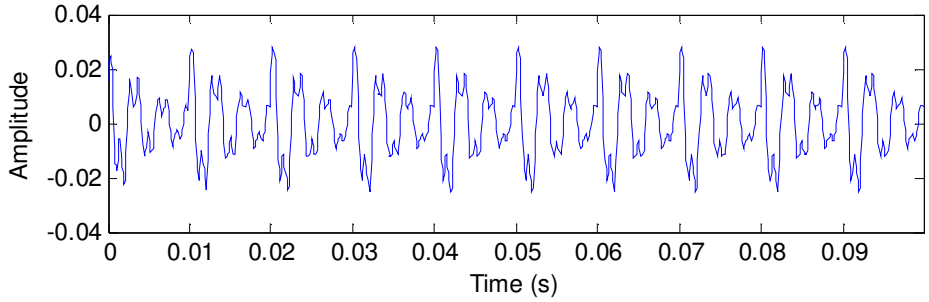
(d)

Figure 5. 1 : *(a) Représentation temporelle de la voyelle /a/ Trajectoires formantiques estimées, superposées au scalogramme correspondant à la voyelle /a/ : (b) utilisant l'ondelette cmor10-1, (c) utilisant fbsp10-1-1, (d) utilisant shan0.1-1.*

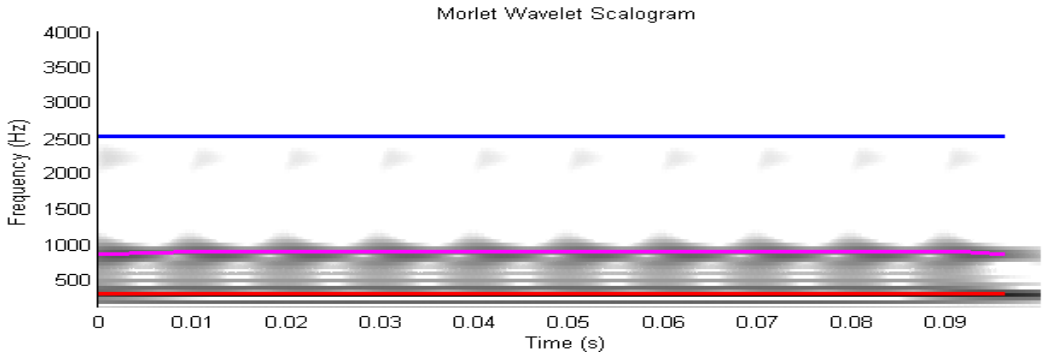
-La voyelle /u/

Les valeurs des trois premiers formants F1, F2 et F3 caractérisant la voyelle /u/ sont respectivement : 300 Hz, 870 Hz et 2240 Hz.

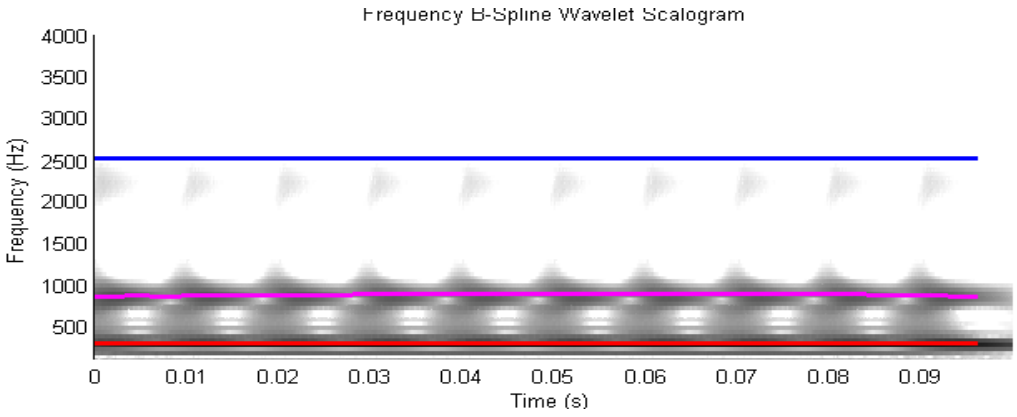
La représentation temporelle de cette voyelle, ainsi que les suivis automatiques des formants basés sur la détection des crêtes d'ondelettes dans le plan temps-fréquence en testant les trois types d'ondelettes sont illustrés par les figures suivantes.



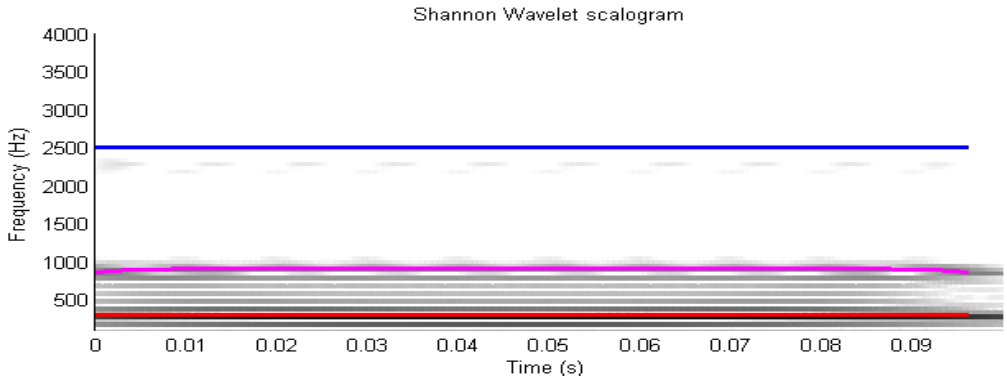
(a)



(b)



(c)



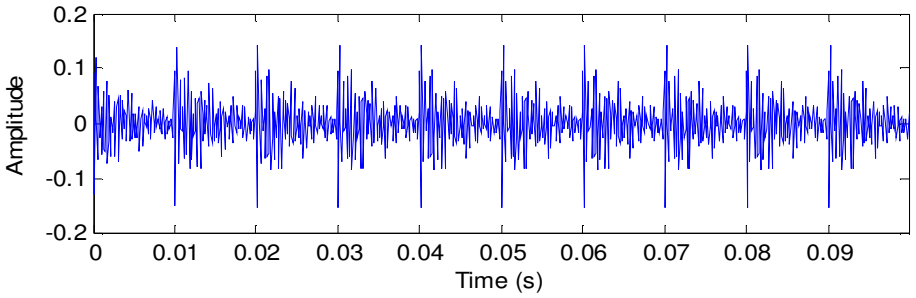
(d)

Figure 5. 2 : *(a) Représentation temporelle de la voyelle /u/ Trajectoires formantiques estimées, superposées au scalogramme correspondant à la voyelle /u/ : (b) utilisant l'ondelette cmor10-1, (c) utilisant fbsp10-1-1, (d) utilisant shan0.1-1.*

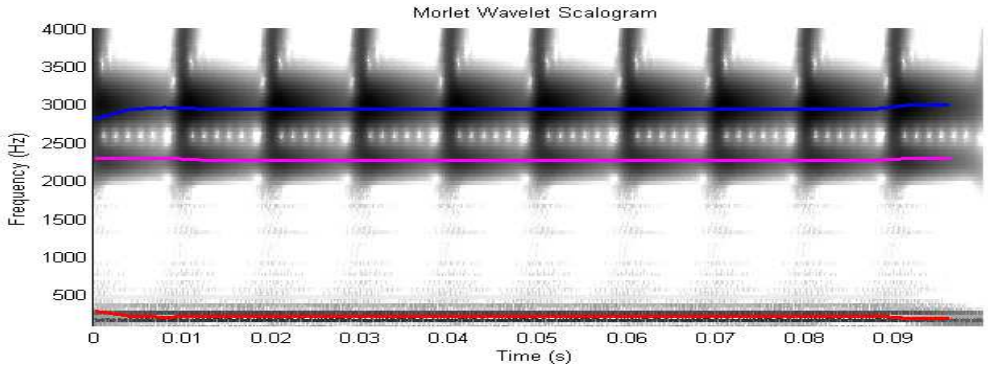
-La voyelle /i/

Les valeurs des trois premiers formants F1, F2 et F3 caractérisant la voyelle /i/ sont respectivement : 270 Hz, 2290 Hz et 3010 Hz.

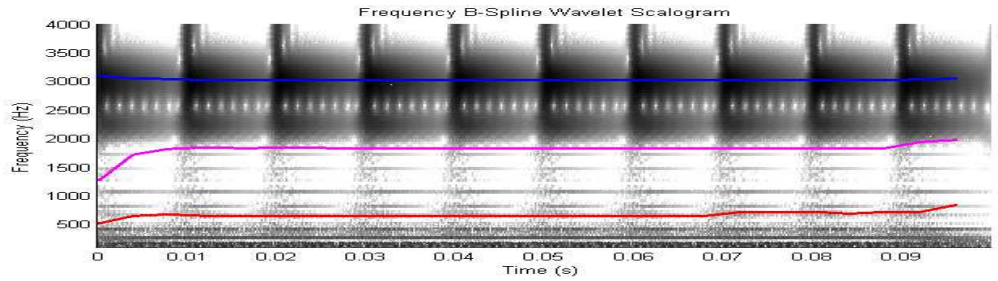
La représentation temporelle de cette voyelle, ainsi que les suivis automatiques des formants basés sur la détection des crêtes d'ondelettes dans le plan temps-fréquence en testant les trois types d'ondelettes sont illustrés par les figures suivantes.



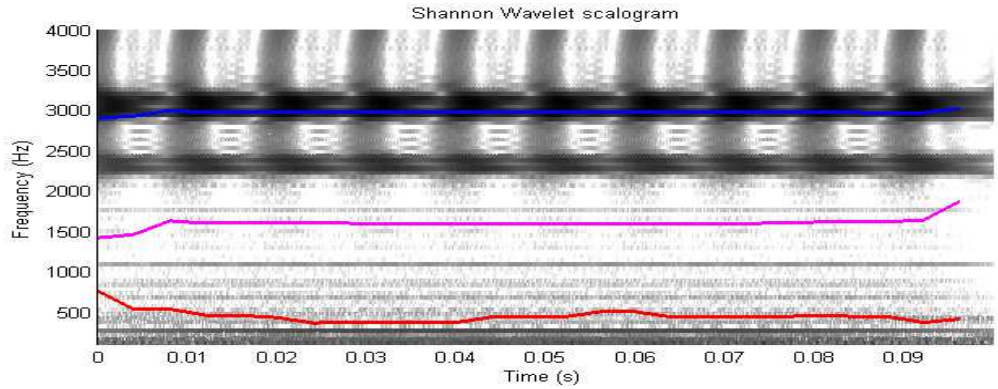
(a)



(b)



(c)



(d)

Figure 5. 3 : *(a) Représentation temporelle de la voyelle /i/ Trajectoires formantiques estimées, superposées au scalogramme correspondant à la voyelle /i/ : (b) utilisant l'ondelette cmor10-1, (c) utilisant fbsp10-1-1, (d) utilisant shan0.1-1.*

➢ **Interprétations des résultats**

Nous remarquons que dans le cas de la voyelle /a/, les résultats fournis par notre algorithme de suivi de formants par détection d'ondelettes (voir Fig.5.1) montrent la capacité de ce dernier à fournir une estimation fiable des trajectoires formantiques. Nous remarquons que les résultats sont bons aussi pour F1 et F2 dans le cas de la voyelle /u/ pour la détection en utilisant les trois ondelettes (voir Fig.5.2) sauf dans le cas de F3 à cause de sa faible énergie au niveau du scalogramme. Dans le cas de la voyelle /i/, (voir Fig.5.3), les résultats sont bons seulement pour la détection avec l'ondelette cmor. Pour les autres ondelettes, l'estimation de F3 est bonne mais ce n'est pas le cas pour F1 et F2 (voir les cas encadrés en rouge dans le tableau 5.1 présenté ci-dessous). Lorsqu'on calcule le scalogramme en utilisant les ondelettes fbsp et shan, nous constatons que le compromis entre la résolution du scalogramme et l'estimation des fréquences des formants est moins pertinent. Les résultats des figures Fig.5.3.c.d montrent en effet que la résolution du scalogramme n'est pas bonne par comparaison au scalogramme calculé avec l'ondelette cmor (voir les Fig.5.3.b).

Il apparaît donc que le suivi de formants utilisant l'ondelette cmor est souvent largement meilleur que celui avec les autres ondelettes et aussi que le suivi de formants et la résolution du scalogramme donnés par les ondelettes cmor et fbsp sont meilleurs qu'avec l'ondelette shan.

Pour confirmer ces observations, nous avons aussi effectué une évaluation quantitative de ces tests sur les voyelles synthétiques en testant les trois ondelettes (cmor, fbsp et shan) tout en calculant la moyenne de différence absolue (Eq.5.1) notée (Diff en Hz) et l'écart-type normalisé en fonction des valeurs de référence (Eq.5.2) notée (σ en %). Ces opérations portent sur les formants estimés et la référence pour chaque formant (F1, F2 et F3) c'est-à-dire les valeurs de formants connues priori avant de synthétiser chaque voyelle (voir le tableau 5.1 ci-dessous).

$$Diff = \frac{1}{N} \times \sum_{p=1}^{N} \left| F_r(p) - F_c(p) \right| \qquad \text{(Eq.5.1)}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{p=1}^{N} \left(\frac{\left| F_r(p) - F_c(p) \right|}{F_r} \right)^2} \qquad \text{(Eq.5.2)}$$

Avec F_c la valeur estimée du formant, F_r la valeur de référence, N le nombre total des fréquences formantiques pour chaque formant et p le compteur de la somme qui va de 1 à N.

		Crêtes d'Ondelette: CMOR			Crêtes d'Ondelette: FBSP			Crêtes d'Ondelette: SHAN		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3
/a/	Diff	27	26	30	42	14	29	36	37	54
	σ	4	4	4	6	3	4	5	5	8
/u/	Diff	0	18	270	0	8	270	8	47	270
	σ	0	6	90	0	3	90	3	16	90
/i/	Diff	39	15	66	407	472	17	190	681	36
	σ	15	6	27	152	181	9	76	253	15

Tableau 5. 1 : *Résultats obtenus sur les voyelles synthétiques /a/, /i/ et /u/*

Les valeurs colorées affichées dans le tableau 5.1 sont les fortes valeurs de différence (supérieures ou égales à 90Hz) entre les valeurs estimées par chaque méthode automatique de suivi et la valeur de référence. Les valeurs en rouge correspondent au calcul de la différence moyenne absolue (Diff en Hz) et les valeurs en bleu correspondent au calcul des écarts type correspondants (σ en %). Les cases encadrées en rouge dans le tableau montrent bien que les erreurs concernent certains cas seulement, ce que confirme les observations visuelles.

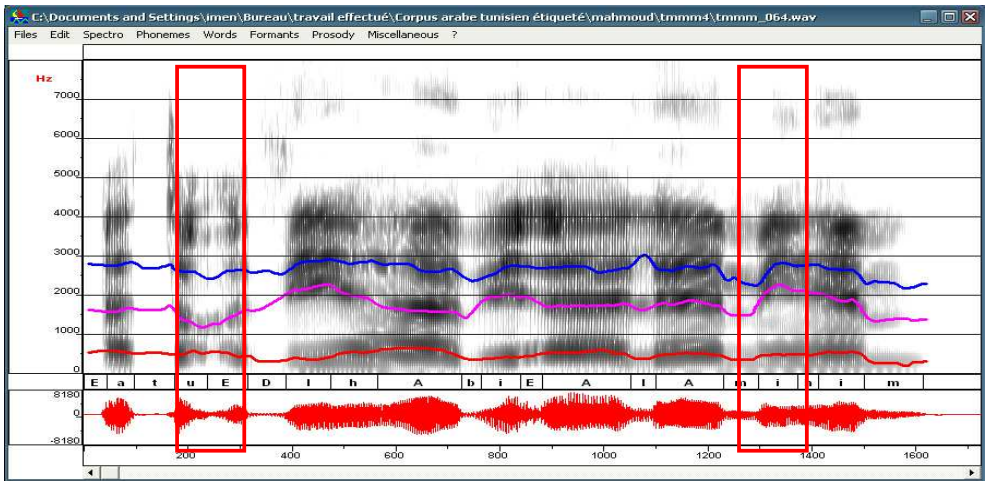
5.3. Tests sur les signaux réels et interprétations

Après l'évaluation de la méthode proposée sur des voyelles synthétiques, nous avons testé notre algorithme de suivi de formants en utilisant le calcul de centre de gravité sur des signaux de la parole naturelle contenant des segments voisés issus de notre base de données étiquetée. Nous avons fait plusieurs tests en testant les trois ondelettes.

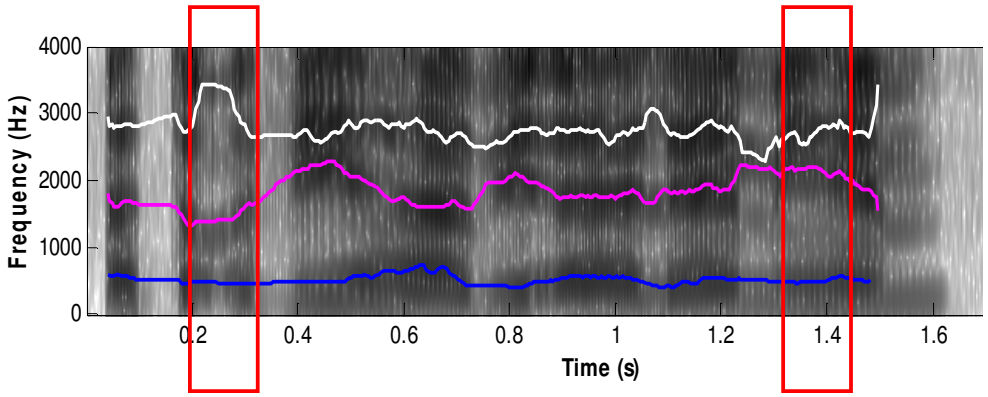
Nous comparons notre nouvelle méthode basée sur les ondelettes avec la méthode automatique de suivi de formants Crêtes de Fourrier en prenant notre étiquetage manuel comme référence.

Dans cette section, nous allons seulement présenter deux exemples de nos tests sur deux phrases en langue Arabe standard issues de notre base (pour lesquels les figures 5.5 et 5.6 montrent le suivi de référence manuel pour F1, F2 et F3) à l'aide de l'algorithme utilisant les crêtes de Fourier et de l'algorithme utilisant les crêtes d'ondelettes. Ces figures vont nous servir pour la comparaison visuelle. Ensuite, dans la section suivante nous allons faire des statistiques pour évaluer les résultats quantitativement.

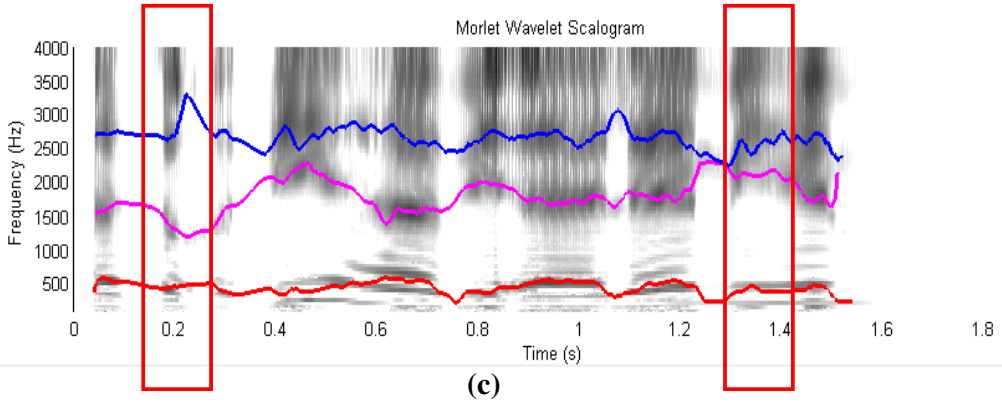
-Exemple1



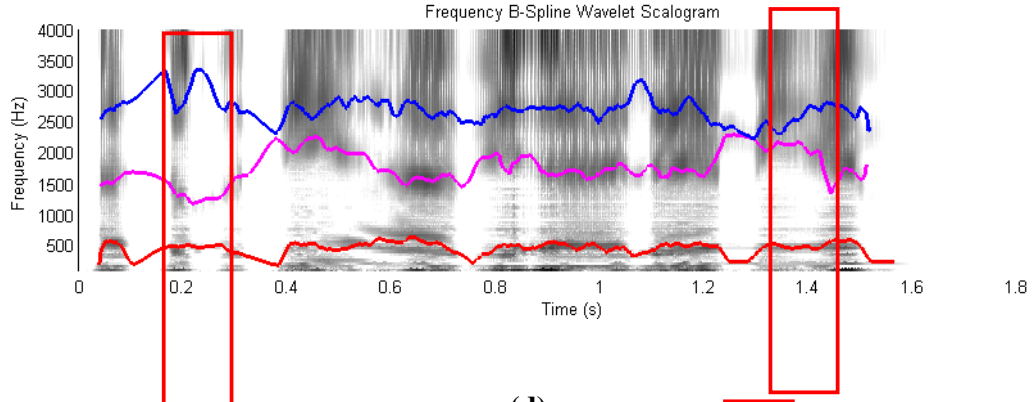
(a)


Frequency (Hz)
4000
3000
2000
1000
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Time (s)

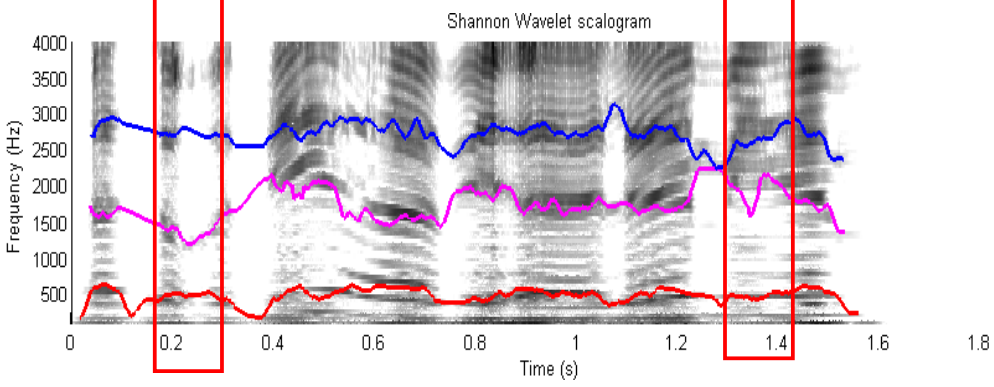
(b)


Morlet Wavelet Scalogram
Frequency (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time (s)

(c)


Frequency B-Spline Wavelet Scalogram
Frequency (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time (s)

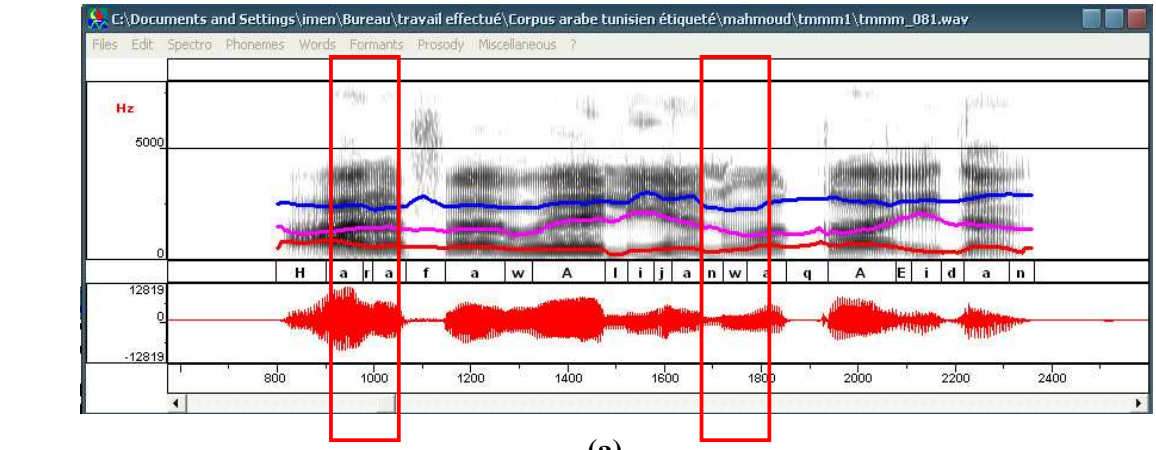
(d)


Shannon Wavelet scalogram
Frequency (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time (s)

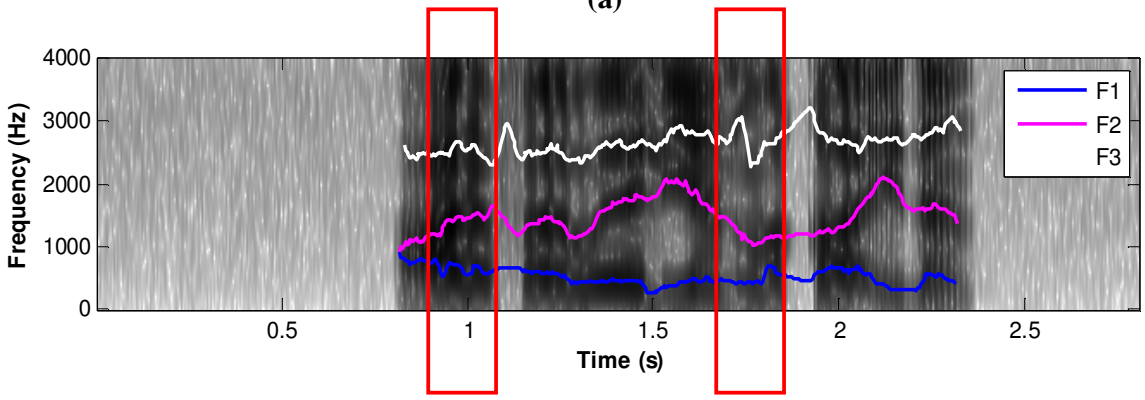
(e)

Figure 5. 4 : *Trajectoires formantiques estimées du signal sonore « أَتُؤْذِيهَا بِالاِمِهِمْ؟» «**ʒatuʒðiːhaː biʒaːlaːmihim ?** » prononcée par le locuteur M1 (a) Etiquetage formantique manuel de référence, (b) Trajectoires formantiques estimées utilisant les crêtes de Fourier et Trajectoires formantiques estimées utilisant les crêtes d'ondelettes : (c) utilisant l'ondelette cmor10-1, (d) utilisant fbsp10-1-1, (e) utilisant shan0.1-1.*

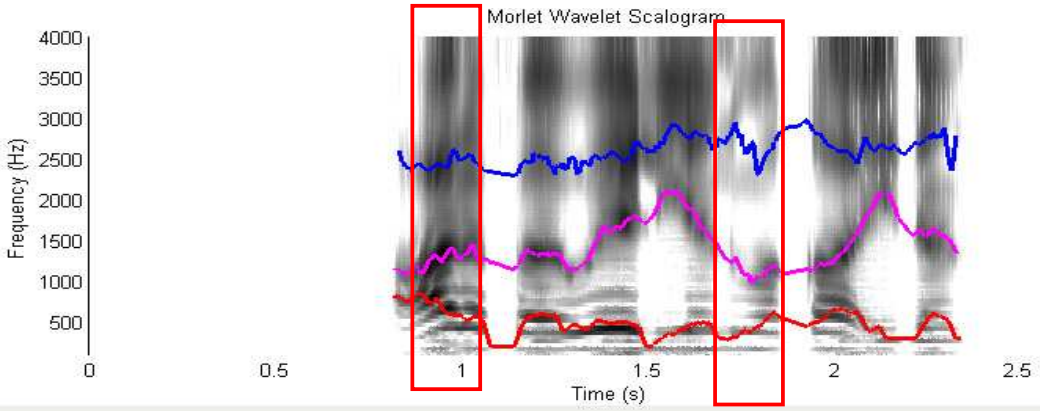
-Exemple 2



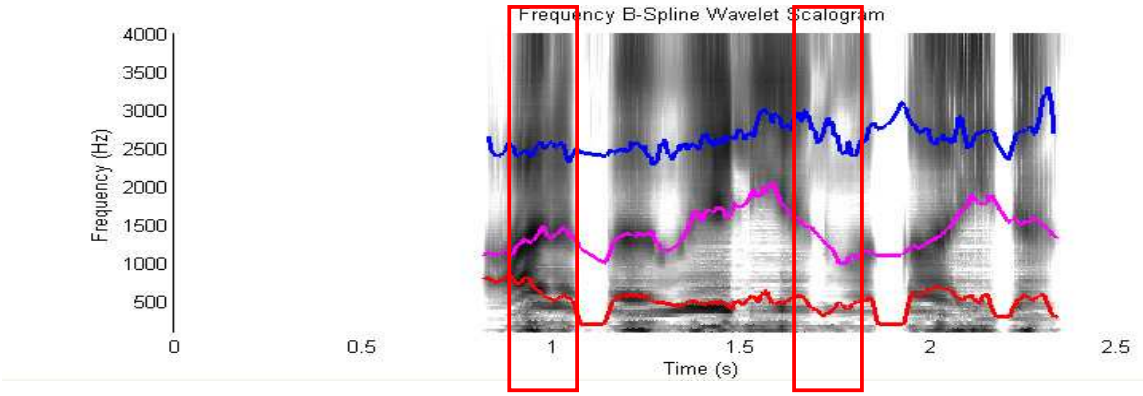
(a)



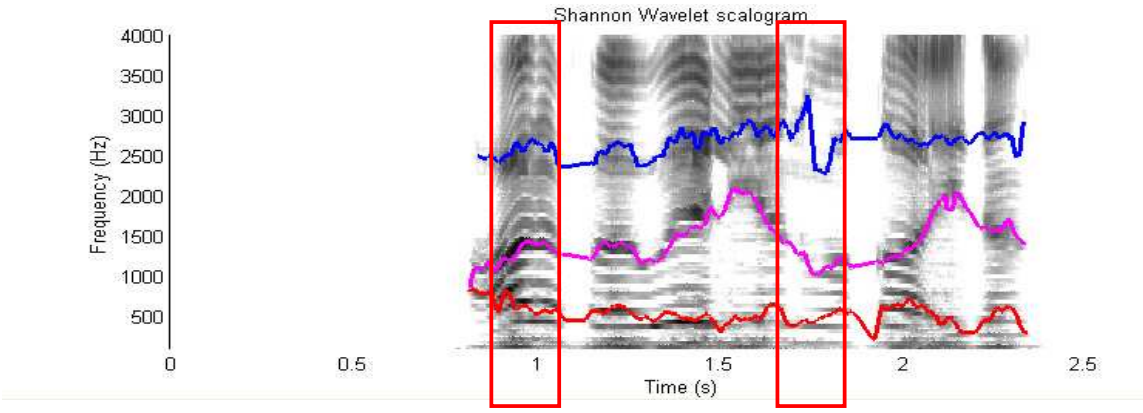
(b)



(c)



(d)



(e)

Figure 5. 5 : *Trajectoires formantiques estimées du signal sonore «*عَرَفَ وَالِيًا وَقَائِدًا*»* «ɛarafa wa:liyan wa qa:3idan» *prononcée par le locuteur M1 (a) Etiquetage formantique manuel de référence, (b) Trajectoires formantiques estimées utilisant les crêtes de Fourier et Trajectoires formantiques estimées utilisant les crêtes d'ondelettes : (c) utilisant l'ondelette cmor10-1, (d) utilisant fbsp10-1-1, (e) utilisant shan0.1-1.*

➢ **Interprétations des résultats**

Dans le cas du 1er exemple (voir Fig.5.4), prononcé par le locuteur M1, les résultats montrent que la méthode proposée trouve bien le bon suivi pour (F1, F2 et F3) et les voyelles /u/ et /i/ par comparaison à l'étiquetage de référence et la méthode par crêtes de Fourier. Les résultats sont bons en utilisant les trois ondelettes pour la première occurrence de la voyelle /u/. Par contre pour /i/ les résultats ne sont bons que pour les ondelettes cmor et fbsp car l'ondelette shan ne donne pas le suivi correct pour F2.

Dans le cas de 2ème exemple (Voir Fig.5.5), prononcé par le locuteur M1, les résultats du suivi sur /ara/ montrent que la méthode proposée présente globalement un suivi presque similaire à celui de l'étiquetage référence, et cela est valable pour les deux ondelettes fbsp et cmor ainsi que la méthode crête de Fourier. En revanche dans le cas de

l'ondelette shan (voir Fig5.5.e) le suivi présente quelques erreurs pour les trois formants probablement à cause de la résolution du scalogramme. Pour la séquence /nwa/, on note que les résultats présentent un bon suivi pour F1 et F2 par rapport à la référence pour toutes les représentations.

A travers ces deux exemples, on constate que les résultats de la méthode proposée pour les ondelettes cmor et fbsp sont globalement presque similaires. Par contre lorsqu'on calcule le scalogramme en utilisant l'ondelette shan, nous constatons que le compromis entre la résolution du scalogramme et l'estimation des fréquences des formants est moins pertinent (même remarque que tout à l'heure avec les voyelles synthétiques). On voit bien dans les figures (Fig.5.4.e et Fig.5.5.e) que les formants ne sont pas bien mis en évidence et qu'il y a une apparition des harmoniques.

Il apparaît donc que le suivi de formants et la résolution des scalogrammes donnés par les ondelettes cmor et fbsp sont meilleurs qu'avec l'ondelette shan même pour les signaux réels.

5.4. Etude et évaluations quantitatives de l'algorithme de suivi de formants utilisant le calcul de centre de gravité

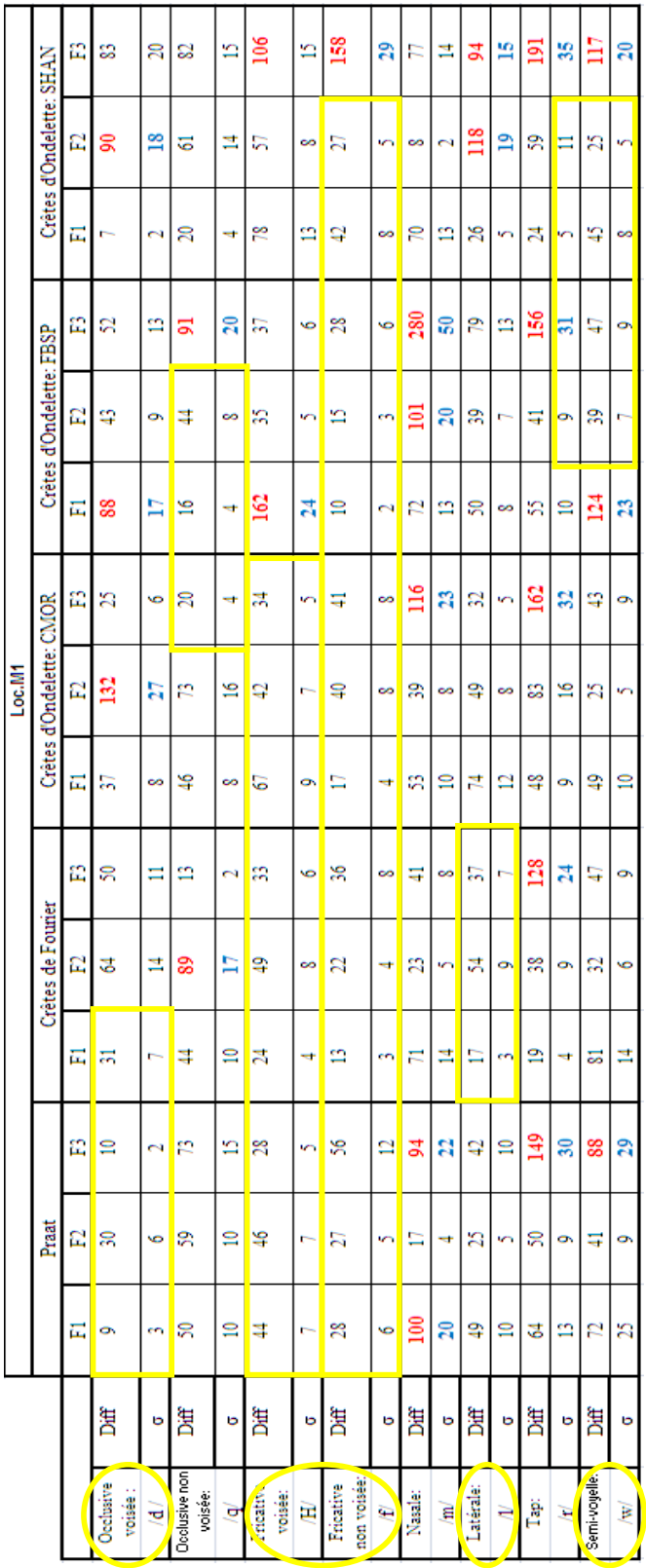
Afin d'évaluer notre nouvel algorithme de suivi de formants par détection des crêtes d'ondelette en utilisant le calcul de centre de gravité, nous avons d'abord fait des tests sur la voyelle /a/ précédée chaque fois d'une consonne. Ensuite, nous avons testé l'algorithme sur les différentes voyelles courtes et longues. Les résultats de suivi ont été ensuite comparés à ceux des méthodes crêtes de Fourier et d'analyse LPC mise en œuvre dans le logiciel (Praat). Finalement, nous avons testé notre algorithme sur la parole continue en faisant des tests sur des phrases et les résultats de suivi ont été comparés à ceux de la méthode par crêtes de Fourier en utilisant le calcul de centre de gravité, de l'analyse LPC combinée à des bancs de filtres de Mustafa Kamran (Mustafa.K 2003) (Mustafa.K 2006) et de l'analyse LPC dans le logiciel Praat. Faute de temps nous n'avons pas pu ajouter aussi la méthode automatique basée sur l'analyse LPC mise en œuvre dans le logiciel Winsnoori pour la comparaison des résultats puisque l'évaluation statistique nous a pris beaucoup de temps et comme on a déjà fait l'étiquetage formantique manuel avec l'outil Winsnoori nous avons préféré de comparer nos résultats avec un autre outil qui est Praat pour enrichir nos travaux.

5.4.1. Etude et évaluations quantitatives sur différentes voyelles

Les tableaux 5.2, 5.3 et 5.4 ci-dessous montrent les résultats obtenus sur la voyelle / a / précédée chaque fois d'une consonne de chaque classe (occlusive voisée, occlusive non voisée, fricative voisée, fricative non voisée, nasale, latéral, tap, semi-voyelle), en calculant la moyenne de différence absolue (Eq.5.1) notée (Diff en Hz) et l'écart-type normalisé en fonction des valeurs de référence (Eq.5.2) notée (σ en %). Ces opérations portent sur les formants estimés et la référence pour chaque formant (F1, F2 et F3). Les combinaisons CV ont été extraites des quatre phrases suivantes :

- « عَرَفَ وَالِيًا وَقَائِدًا» « **ʕarafa wa:liyan wa qa:ʔidan** » (Il connait un gouverneur et un commandant)
- « هِيَ هُنَا لَقَدْ آبَتْ» « **Hiya huna: laqad ʒa:bat** ». (Elle était ici et elle était pieuse)
- « لَقَدْ كَانَ مُسَالِمًا وَقُتِلَ. » «**Laqad ka:na musa:liman wa qutila** » (C'était un pacifiste et il a été tué)
- «قَادَ الجَيْشَ.ا» «***qa:da ljayŝa***» (Il a commandé l'armée)

Ces phrases ont été prononcées par les cinq locuteurs masculins: M1, M2, M3, M4 et M5. Le premier tableau 5.2 du locuteur M1 donne les résultats quantitatifs de la nouvelle méthode basée sur la détection des crêtes d'ondelette en testant les trois types d'ondelettes (cmor, fbsp et shan) et ceux des méthodes Fourier et LPC du logiciel (Praat). Les résultats présentés par les tableaux 5.3 et 5.4 des locuteurs M2, M3, M4 et M5 ne portent que sur la voyelle /a/ en testant la nouvelle méthode avec l'ondelette cmor ainsi que les deux autres méthodes de comparaison.



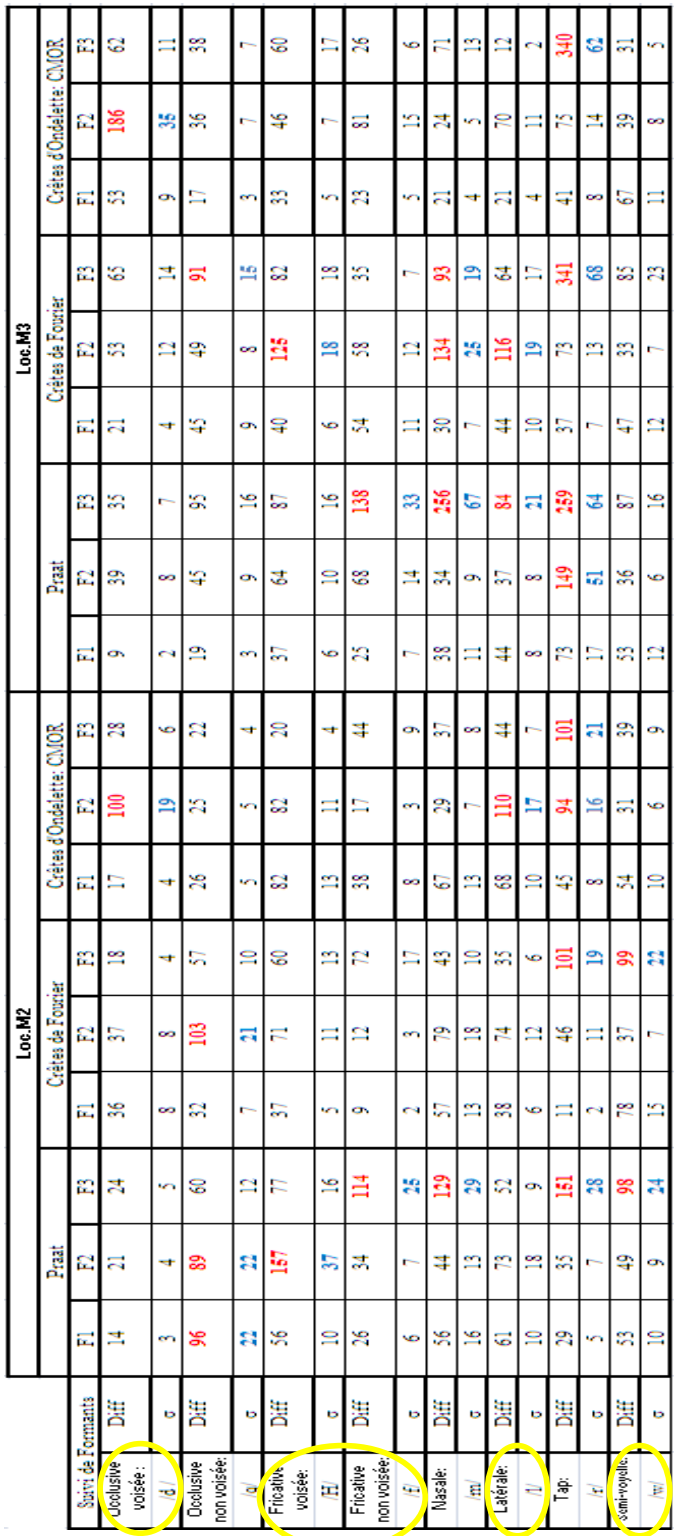
Loc.M1		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR			Crêtes d'Ondelette: FBSP			Crêtes d'Ondelette: SHAN		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Occlusive voisée : /d/	Diff	9	30	10	31	64	50	37	132	25	88	43	52	7	90	83
	σ	3	6	2	7	14	11	8	27	6	17	9	13	2	18	20
Occlusive non voisée: /q/	Diff	50	59	73	44	89	13	46	73	20	16	44	91	20	61	82
	σ	10	10	15	10	17	2	8	16	4	4	8	20	4	14	15
Fricative voisée: /H/	Diff	44	46	28	24	49	33	67	42	34	162	35	37	78	57	106
	σ	7	7	5	4	8	6	9	7	5	24	5	6	13	8	15
Fricative non voisée: /f/	Diff	28	27	56	13	22	36	17	40	41	10	15	28	42	27	158
	σ	6	5	12	3	4	8	4	8	8	2	3	6	8	5	29
Nasale: /m/	Diff	100	17	94	71	23	41	53	39	116	72	101	280	70	8	77
	σ	20	4	22	14	5	8	10	8	23	13	20	50	13	2	14
Latérale: /l/	Diff	49	25	42	17	54	37	74	49	32	50	39	79	26	118	94
	σ	10	5	10	3	9	7	12	8	5	8	7	13	5	19	15
Tap: /r/	Diff	64	50	149	19	38	128	48	83	162	55	41	156	24	59	191
	σ	13	9	30	4	9	24	9	16	32	10	9	31	5	11	35
Semi-voyelle: /w/	Diff	72	41	88	81	32	47	49	25	43	124	39	47	45	25	117
	σ	25	9	29	14	6	9	10	5	9	23	7	9	8	5	20

Tableau 5. 2 : *Résultats obtenus sur la voyelle/a/ précédée de chaque type de consonne sur des signaux prononcés par le locuteur masculin M1*

- **Interprétations des résultats**

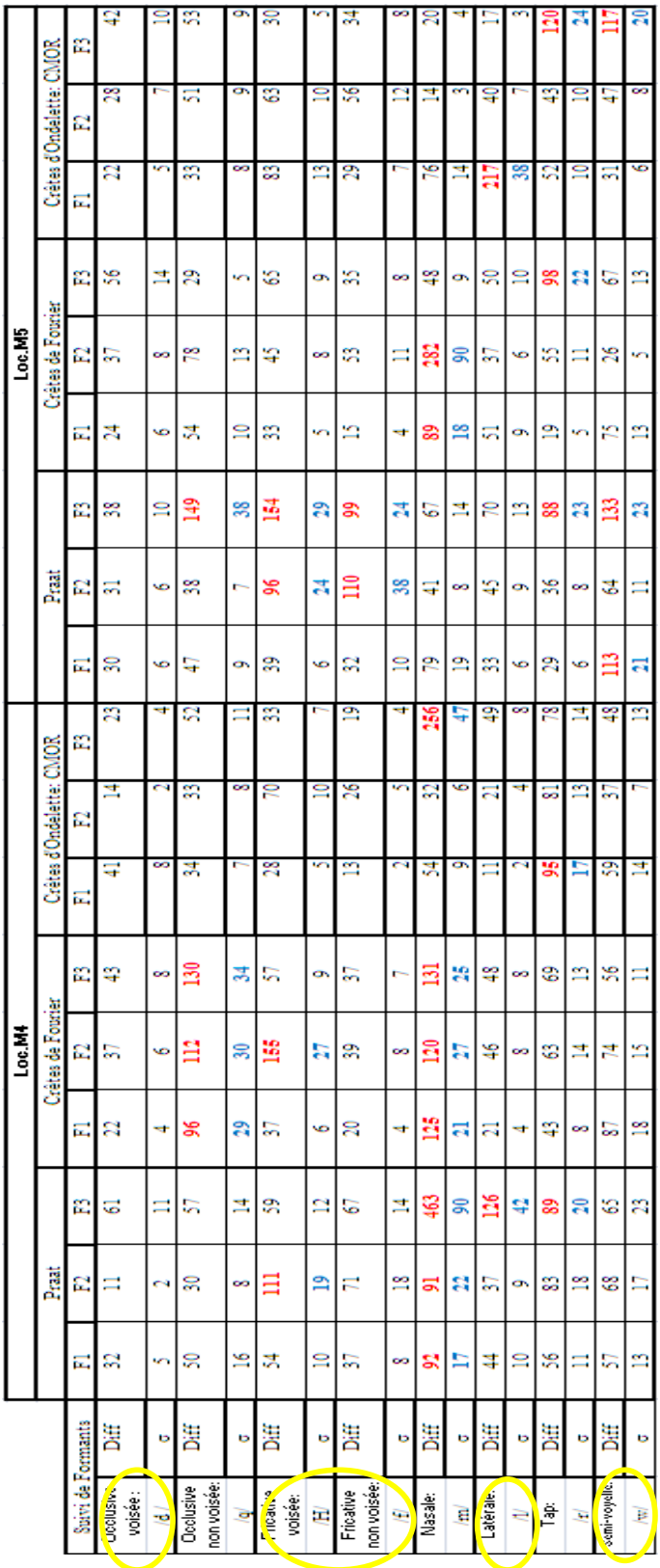
Les valeurs colorées affichées dans le tableau 5.2 sont des valeurs d'erreurs élevées supérieures ou égales à 90Hz entre les valeurs estimées par chaque méthode automatique de suivi et la valeur référence étiquetée manuellement. Les valeurs en rouges correspondent au calcul de la différence moyenne absolue (Diff en Hz) et les valeurs en bleu au calcul des écarts type correspondants (σ en %).

La comparaison des résultats entre les trois ondelettes montre que les résultats obtenus avec cmor sont légèrement meilleurs que ceux des autres ondelettes. L'utilisation des trois ondelettes peut donner à un bon suivi de formants puisque les valeurs de la différence moyenne absolue et de l'écart type sont très faibles, (voir tableau 5.2), ce qui signifie que le suivi de formants est proche ou presque conforme à l'étiquetage formantique manuel. Les cases encadrées en jaune correspondent à cette situation. Par exemple dans le cas où la voyelle /a/ est précédée par la consonne fricative non voisée /f/ on obtient des erreurs avec l'ondelette cmor pour F1 de 17Hz (σ=4%), pour F2 de 40Hz (σ=8%) et pour F3 de 41Hz (σ=8%) de même pour les autres ondelettes par exemple pour le suivi fait par fbsp on a pour F1 de 10Hz (σ=2%), pour F2 de 15Hz (σ=3%) et pour F3 de 28Hz (σ=6%). Comme on le voit dans le tableau, il y a de nombreux cas pour lesquels les écarts type sont très faibles (inférieurs à 10Hz) ce qui signifie que le suivi de formants automatique est très proche de suivi référence. Malgré tout il apparaît que l'ondelette cmor conduit à de meilleurs résultats par rapport aux autres ondelettes fbsp (7 valeurs en rouge) et shan (6 valeurs en rouge). Les résultats montrent que le suivi est globalement bien pour les trois formants de la voyelle courte /a/ quelle que soit la nature de la consonne. Cela est vrai pour les trois méthodes Praat, Fourier et ondelette, puisque les valeurs de la différence moyenne et de l'écart type sont faibles (voir quelques cases du tableau 5.2 qui sont encadrés en jaune pour certains cas). Il y a plutôt des erreurs localisées (valeurs en rouge) pour F3 dans le cas où la consonne précédente est /r/ (voir tableau 5.2).



Suivi de Formants		Loc.M2						Loc.M3					
		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CWOR			Praat			Crêtes de Fourier			Crêtes d'Ondelette: CWOR		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Occlusive voisée: /d/	Diff	14	21	24	36	37	18	17	100	28	9	39	35	21	33	65	53	186	62
	σ	3	4	5	8	8	4	4	19	6	2	8	7	4	12	14	9	35	11
Occlusive non voisée: /t/	Diff	96	89	60	32	103	57	26	25	22	19	45	95	45	49	91	17	36	38
	σ	22	22	12	7	21	10	5	5	4	3	9	16	9	8	15	3	7	7
Fricative voisée: /f/	Diff	56	157	77	37	71	60	82	82	20	37	64	87	40	125	82	33	46	60
	σ	10	37	16	5	11	13	13	11	4	6	10	16	6	18	18	5	7	17
Fricative non voisée: /s/	Diff	26	34	114	9	12	72	38	17	44	25	68	138	54	58	35	23	81	26
	σ	6	7	25	2	3	17	8	3	9	7	14	33	11	12	7	5	15	6
Nasale: /m/	Diff	56	44	129	57	79	43	67	29	37	38	34	256	30	134	93	21	24	71
	σ	16	13	29	13	18	10	13	7	8	11	9	67	7	25	19	4	5	13
Latérale: /l/	Diff	61	73	52	38	74	35	68	110	44	44	37	84	44	116	64	21	70	12
	σ	10	18	9	6	12	6	10	17	7	8	8	21	10	19	17	4	11	2
Tap: /r/	Diff	29	35	151	11	46	101	45	94	101	73	149	259	37	73	341	41	75	340
	σ	5	7	28	2	11	19	8	16	21	17	51	64	7	13	68	8	14	62
Semi-voyelle: /w/	Diff	53	49	98	78	37	99	54	31	39	53	36	87	47	33	85	67	39	31
	σ	10	9	24	15	7	22	10	6	9	12	6	16	12	7	23	11	8	5

Tableau 5. 3 : *Résultats obtenus sur la voyelle/a/ précédée par chaque type de consonne traités sur des signaux prononcés par les locuteurs masculins M2 et M3*



		Loc.M4								Loc.M5							
		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR	
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Occlusive voisée : /d/	Diff	32	11	61	22	37	43	41	14	23	30	31	38	24	37	56	22	28	42
	σ	5	2	11	4	6	8	8	2	4	6	6	10	6	8	14	5	7	10
Occlusive non voisée: /p/	Diff	50	30	57	96	112	130	34	33	52	47	38	149	54	78	29	33	51	53
	σ	16	8	14	29	30	34	7	8	11	9	7	38	10	13	5	8	9	9
Fricative voisée: /ES/	Diff	54	111	59	37	165	57	28	70	33	39	96	164	33	45	65	83	63	30
	σ	10	19	12	6	27	9	5	10	7	6	24	29	5	8	9	13	10	5
Fricative non voisée: /f/	Diff	37	71	67	20	39	37	13	26	19	32	110	99	15	53	35	29	56	34
	σ	8	18	14	4	8	7	2	5	4	10	38	24	4	11	8	7	12	8
Nasale: /m/	Diff	92	91	463	125	120	131	54	32	256	79	41	67	89	282	48	76	14	20
	σ	17	22	90	21	27	25	9	6	47	19	8	14	18	90	9	14	3	4
Latérale: /l/	Diff	44	37	126	21	46	48	11	21	49	33	45	70	51	37	50	217	40	17
	σ	10	9	42	4	8	8	2	4	8	6	9	13	9	6	10	38	7	3
Tap: /r/	Diff	56	83	89	43	63	69	95	81	78	29	36	88	19	55	98	52	43	120
	σ	11	18	20	8	14	13	17	13	14	6	8	23	5	11	22	10	10	24
Semi-voyelle: /w/	Diff	57	68	65	87	74	56	59	37	48	113	64	133	75	26	67	31	47	117
	σ	13	17	23	18	15	11	14	7	13	21	11	23	13	5	13	6	8	20

Tableau 5. 4 : *Résultats obtenus sur la voyelle/a/ précédée par chaque type de consonne traités sur des signaux prononcés par les locuteurs masculins M4 et M5*

- **Interprétations des résultats**

Les résultats présentés dans les tableaux 5.2, 5.3 et 5.4 ont été obtenus sur la voyelle /a/ précédée par chaque type de consonne. Les occurrences de cette voyelle sont extraites des phrases prononcées par cinq locuteurs masculins M1, M2, M3, M4 et M5. Le patron formantique de la voyelle est pris au milieu de façon à supprimer les transitions des consonnes adjacentes à cette voyelle à droite et à gauche pour éviter l'influence de la coarticulation avec les consonnes adjacentes.

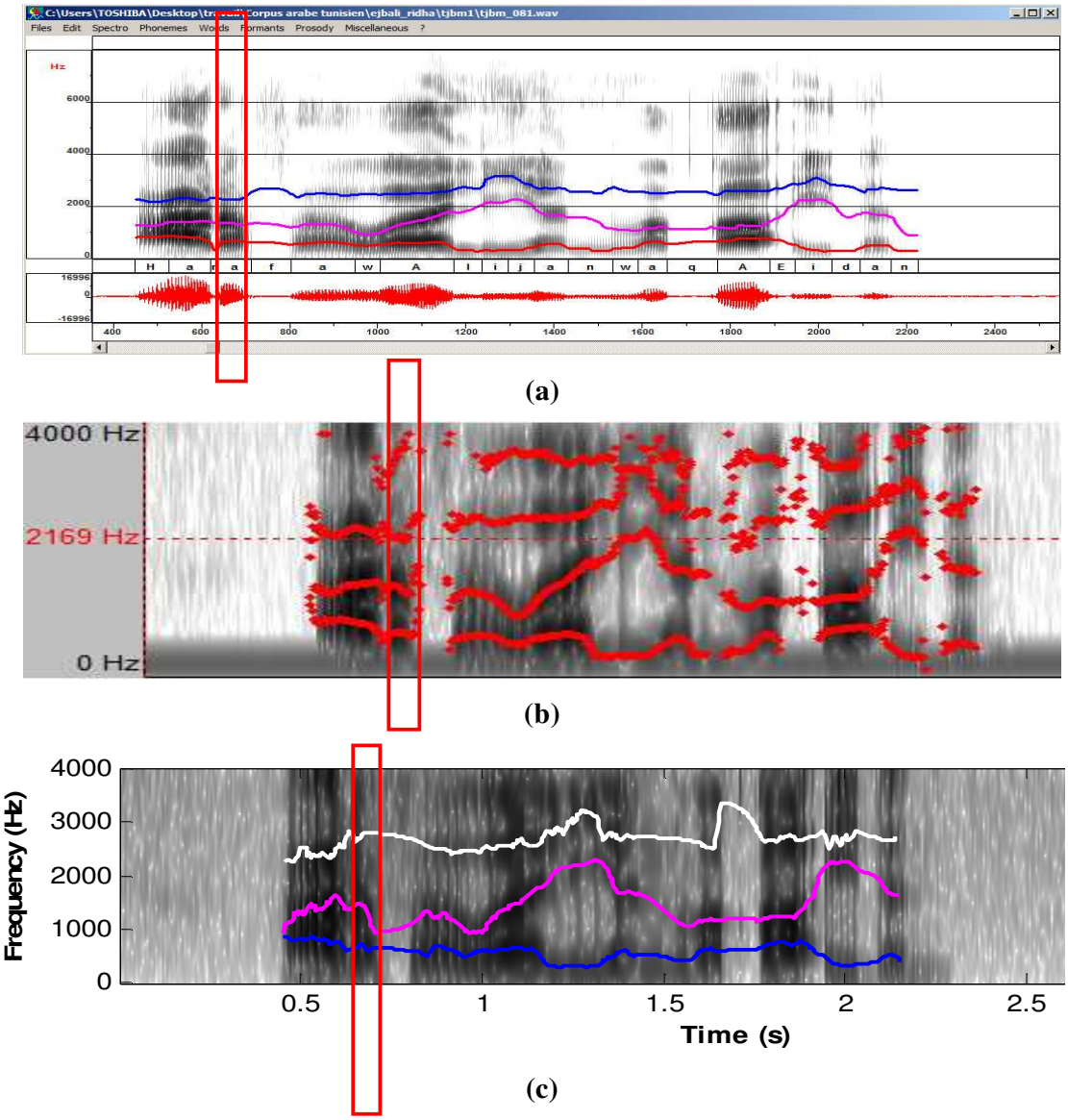
D'après les tableaux (5.2 5.3 et 5.4), on constate que le nombre d'erreurs (valeurs en rouge de la différence moyenne élevé) est globalement plus faible pour le suivi fait par la méthode ondelette avec cmor que pour les autres méthodes Praat et Fourier. Cette méthode donne donc un suivi de formants (F1, F2 et F3) pertinent et plus proche de suivi référence. Les résultats des méthodes Fourier et ondelette sont très proches dans certains cas puisque toutes les deux présentent moins d'erreurs que la méthode Praat. Par exemple dans le cas du tableau 5.2 du locuteur M1, on note que dans le cas de la voyelle /a/ précédée par la consonne /H/, il y a des valeurs de différence moyenne et d'écart type qui sont très proches pour F2 et F3 par la méthode Fourier (respectivement (Diff=49Hz, σ=8%) et (Diff=33Hz, σ=6%)) comme avec la méthode ondelette (respectivement (Diff=42Hz, σ=7%) pour F2 et (Diff=34Hz, σ=5%) pour F3.

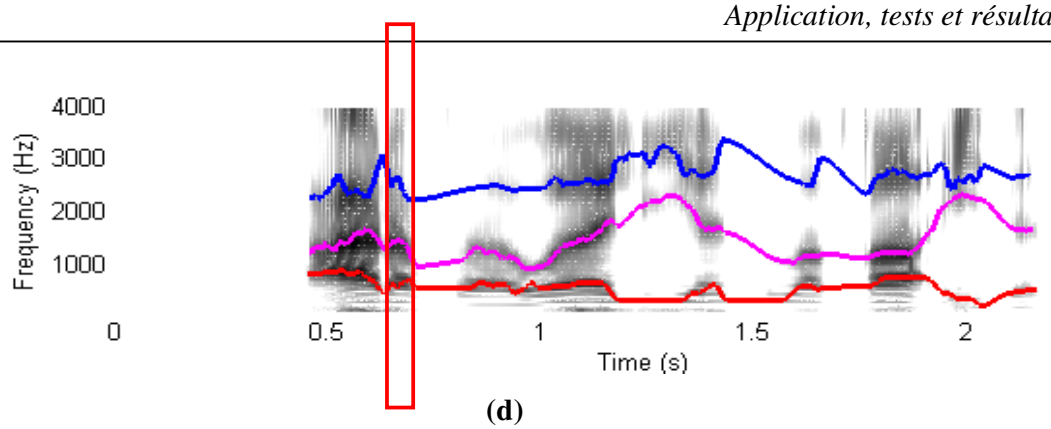
D'après les résultats du tableau 5.5 de locuteur M5, on constate que les suivi des trois formants dans le cas de la voyelle /a/ précédée par la consonne /d/ par les deux méthodes Fourier et ondelette sont très proches. Ces remarques sont valables pour plusieurs autres exemples, les valeurs de Diff et σ des deux méthodes sont bien proches et parfois identiques.

D'après les résultats des valeurs de Diff et σ, on note qu'on obtient un bon suivi pratiquement par les trois méthodes et on a vérifié cela pour les cinq locuteurs. Le suivi des trois formants est pertinent pour la plupart des cas surtout dans les cas où la voyelle /a/ est précédée par les consonnes /d/, /H/, /f/, /l/ et /w/ (voir les cas encerclés en jaune dans les tableaux 5.2, 5.3 et 5.4) car il n'y a pratiquement pas d'erreurs dans ces cas (pas de valeurs en rouges). Par contre dans le cas où la voyelle est précédée par la consonne Tap /r/, on remarque que le suivi est bon pour F1 et F2 par les trois méthodes et pour les cinq locuteurs mais on observe des erreurs (valeurs en rouge) concentrées au niveau de F3 et pour expliquer un peu ces erreurs, on a eu recours à une comparaison visuelle. D'après la figure 5.6, on remarque bien à travers les résultats de suivi des trois méthodes ainsi que la référence que l'énergie de formant F3 de la voyelle /a/ est très faible et à peine visible ce qui explique le faux suivi (suivi

de faux pics) que présente les trois méthodes à cause de l'absence des fréquences candidates correspondantes. Cette faible énergie s'explique peut être par le fait que la voyelle /a/ est précédée par la consonne tap /r/ (très brève) et suivi par la fricative /f/ et cela est dû à l'effet de coarticulation. Ces résultats sont vérifiés pour tous les locuteurs, puisque dans ce cas là il s'agit de la même phrase prononcée.

De même dans les cas où la voyelle /a/ est précédée par les consonnes /m/ et /q/, le suivi est pertinent pour tous les locuteurs avec la méthode par ondelette mais entaché de quelques erreurs pour les deux autres méthodes, les erreurs apparaissent soit pour F1, F2 ou F3.



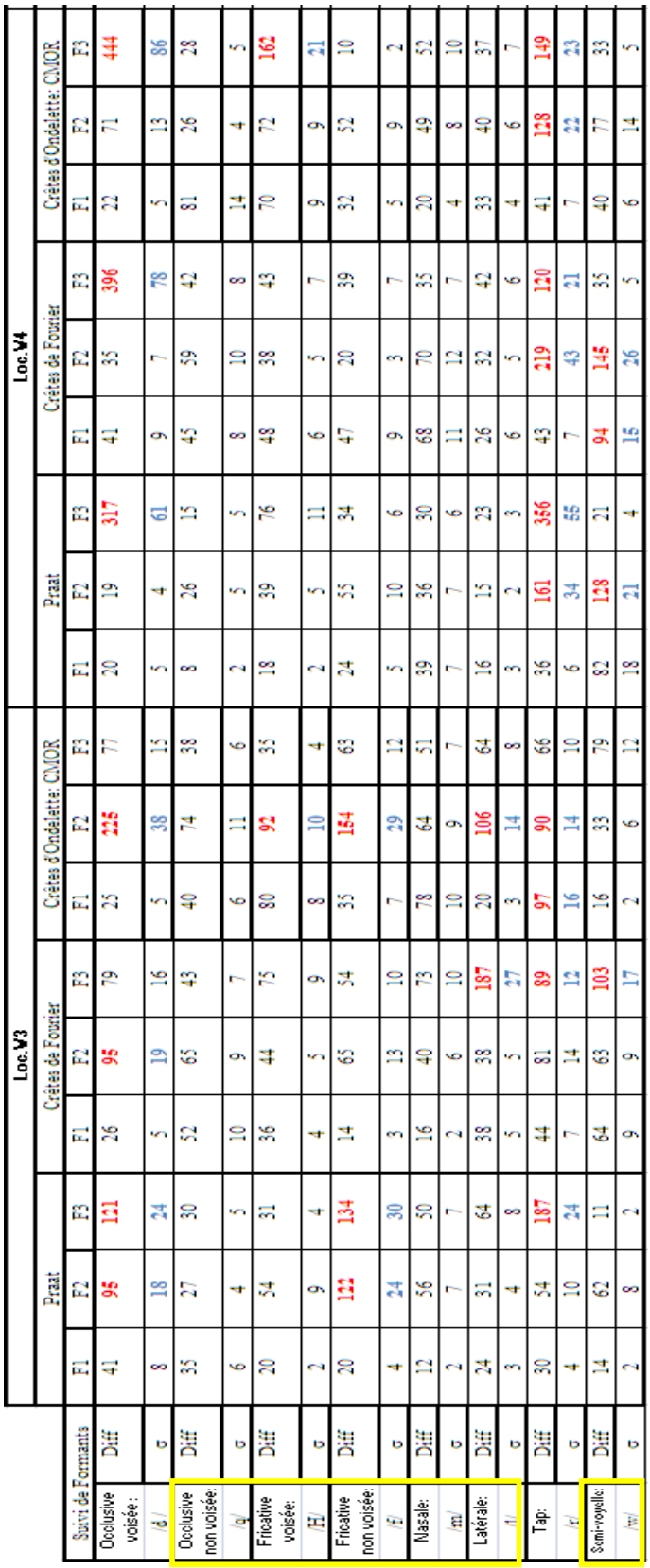


(d)

Figure 5. 6 : *Trajectoires formantiques estimées du signal sonore «*عَرَفَ وَالِيًا وَقَائِدًا*»* «ɛarafa wa:liyan wa qa:3idan*» prononcée par le locuteur M3 (a) Etiquetage formantique manuel de référence,(b)Trajectoires formantiques estimées utilisant Praat (c)Trajectoires formantiques estimées utilisant les crêtes de Fourier et (d) Trajectoires formantiques estimées utilisant les crêtes d'ondelette cmor10-1*

Les tableaux 5.5 et 5.6 ci-dessous montrent les résultats obtenus sur la voyelle / a / précédée chaque fois d'une consonne de chaque classe en calculant la moyenne de différence absolue (Eq.5.1) notée (Diff en Hz) et l'écart-type normalisé en fonction des valeurs de référence (Eq.5.2) notée (σ en %) pour chaque formant (F1, F2 et F3) et pour trois locutrices (W3,W4,W5). Les combinaisons CV ont été extraites des quatre phrases citées précédemment comme pour les locuteurs masculins.

Ces tableaux 5.5 et 5.6 affichent les résultats quantitatifs de la nouvelle méthode basée sur la détection des crêtes d'ondelettes avec cmor et ceux des méthodes Fourier et LPC via le logiciel (Praat).



Loc.W3									Loc.W4								
		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR			Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Occlusive voisée : /d/	Diff	41	95	121	26	95	79	25	225	77	20	19	317	41	35	396	22	71	444
	σ	8	18	24	5	19	16	5	38	15	5	4	61	9	7	78	5	13	86
Occlusive non voisée: /g/	Diff	35	27	30	52	65	43	40	74	38	8	26	15	45	59	42	81	26	28
	σ	6	4	5	10	9	7	6	11	6	2	5	5	8	10	8	14	4	5
Fricative voisée: /Z/	Diff	20	54	31	36	44	75	80	92	35	18	39	76	48	38	43	70	72	162
	σ	2	9	4	4	5	9	8	10	4	2	5	11	6	5	7	9	9	21
Fricative non voisée: /f/	Diff	20	122	134	14	65	54	35	154	63	24	55	34	47	20	39	32	52	10
	σ	4	24	30	3	13	10	7	29	12	5	10	6	9	3	7	5	9	2
Nasale: /m/	Diff	12	56	50	16	40	73	78	64	51	39	36	30	68	70	35	20	49	52
	σ	2	7	7	2	6	10	10	9	7	7	7	6	11	12	7	4	8	10
Latérale: /l/	Diff	24	31	64	38	38	187	20	106	64	16	15	23	26	32	42	33	40	37
	σ	3	4	8	5	5	27	3	14	8	3	2	3	6	5	6	4	6	7
Tap: /r/	Diff	30	54	187	44	81	89	97	90	66	36	161	356	43	219	120	41	128	149
	σ	4	10	24	7	14	12	16	14	10	6	34	55	7	43	21	7	22	23
Semi-voyelle: /w/	Diff	14	62	11	64	63	103	16	33	79	82	128	21	94	145	35	40	77	33
	σ	2	8	2	9	9	17	2	6	12	18	21	4	15	26	5	6	14	5

Tableau 5. 5 : *Résultats obtenus sur la voyelle/a/ précédée par chaque type de consonne traités sur des signaux prononcés par les locutrices W3 et W4*

Loc.W5		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3
Occlusive voisée : /d/	Diff	14	11	141	8	18	151	24	55	176
	σ	2	2	23	2	3	27	4	9	31
Occlusive non voisée: /q/	Diff	30	46	136	49	107	77	33	142	70
	σ	5	7	29	7	14	12	5	18	12
Fricative voisée: /H/	Diff	12	20	66	99	60	68	86	19	183
	σ	2	3	13	12	7	9	10	2	23
Fricative non voisée: /f/	Diff	47	57	130	42	33	54	29	35	48
	σ	9	11	25	8	6	12	5	6	8
Nasale: /m/	Diff	16	73	16	47	89	31	31	35	31
	σ	3	12	3	7	14	5	5	6	5
Latérale: /l/	Diff	55	71	47	24	144	48	58	106	44
	σ	8	12	7	4	22	7	8	17	7
Tap: /r/	Diff	29	19	89	26	17	70	26	58	161
	σ	5	3	17	5	3	10	5	10	28
Semi-voyelle: /w/	Diff	25	36	139	21	45	71	17	19	76
	σ	4	6	24	4	7	15	2	3	16

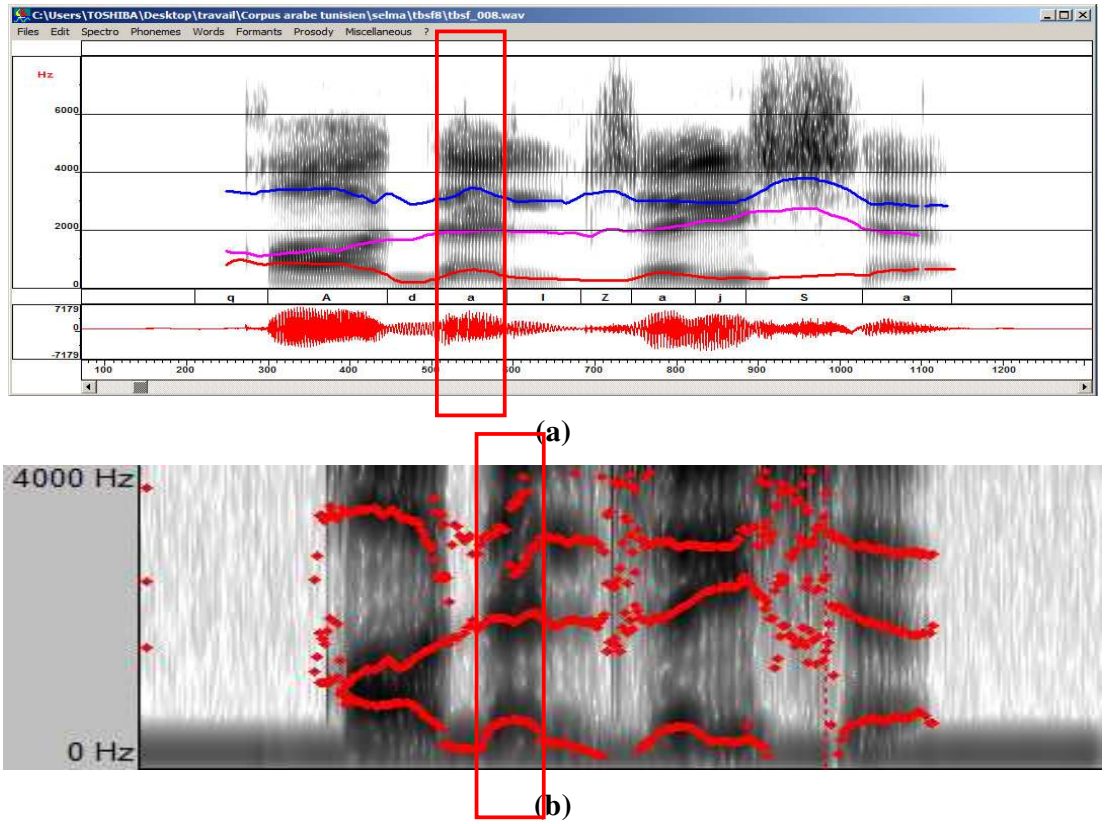
Tableau 5. 6 : *Résultats obtenus sur la voyelle/a/ précédée par chaque type de consonne traités sur des signaux prononcés par la locutrice W5*

➢ **Interprétations des résultats**

D'après les résultats des valeurs des erreurs présentés dans les deux tableaux 5.5 et 5.6, on note bien qu'on a globalement un bon suivi de formants pour les trois méthodes et pour les trois locutrices W3, W4 et W5. Le suivi des trois formants est pertinent pour la plupart des cas surtout dans les cas où la voyelle /a/ est précédée par les consonnes /q/, /H/, /f/, /m/, /l/ et /w/. Il ya seulement quelques erreurs (valeurs en rouge) localisées parfois en F2 dans par la méthode crêtes d'ondelette dans le cas où la voyelle /a/ est précédée par la consonne latérale /l/ et des erreurs parfois en F3 dans le cas où la voyelle /a/ est précédée par la consonne fricative voisée /H/. D'autres erreurs apparaissent localement sur F1, F2 ou F3

par la méthode crêtes de Fourier dans le cas où la voyelle /a/ est précédée par la semi-voyelle /w/. Certaines erreurs concernant F3 sont commises soit par la méthode LPC quand la voyelle est précédée par la fricative non voisée /f/.

Par contre dans le cas où la voyelle est précédée par les consonnes /r/ et /d/, on remarque qu'on a toujours un bon suivi de F1 pour les trois méthodes même s'il y a parfois certaines erreurs locales surtout sur F3 et parfois sur F2. Cette observation vaut pour les trois méthodes et les trois locutrices. Pour expliquer un peu ces erreurs, on a pris l'exemple de la voyelle /a/ précédée par la consonne /d/ et on a eu recours à une comparaison visuelle. D'après les résultats présentée par la figure 5.7, on note que le suivi de formant F3 fait par les trois méthodes dans ce cas là n'est pas conforme au suivi référence ce qui confirme les erreurs concentrées sur F3 pour la locutrice W4 (voir tableau 5.5) et pour les autres locutrices. Ces erreurs peuvent être expliquées par le fait que la consonne /l/ qui suit la voyelle /a/ porte la voyelle muette (soukoun) et à cause de ça et de point de vue énergie, on voit bien que l'énergie de F3 n'est pas bien claire et raffinée comme dans le cas des formants F1 et F2, donc on a tendance de suivre de faux pics dû à l'effet de coarticulation.



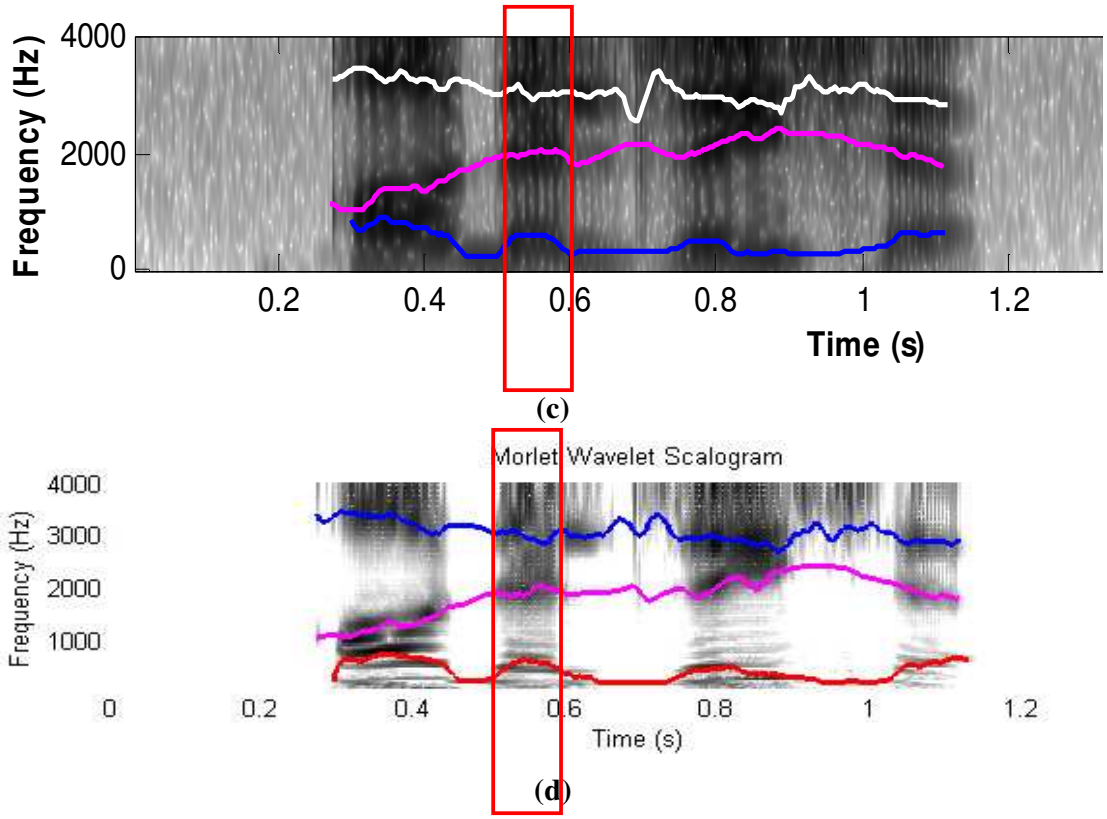
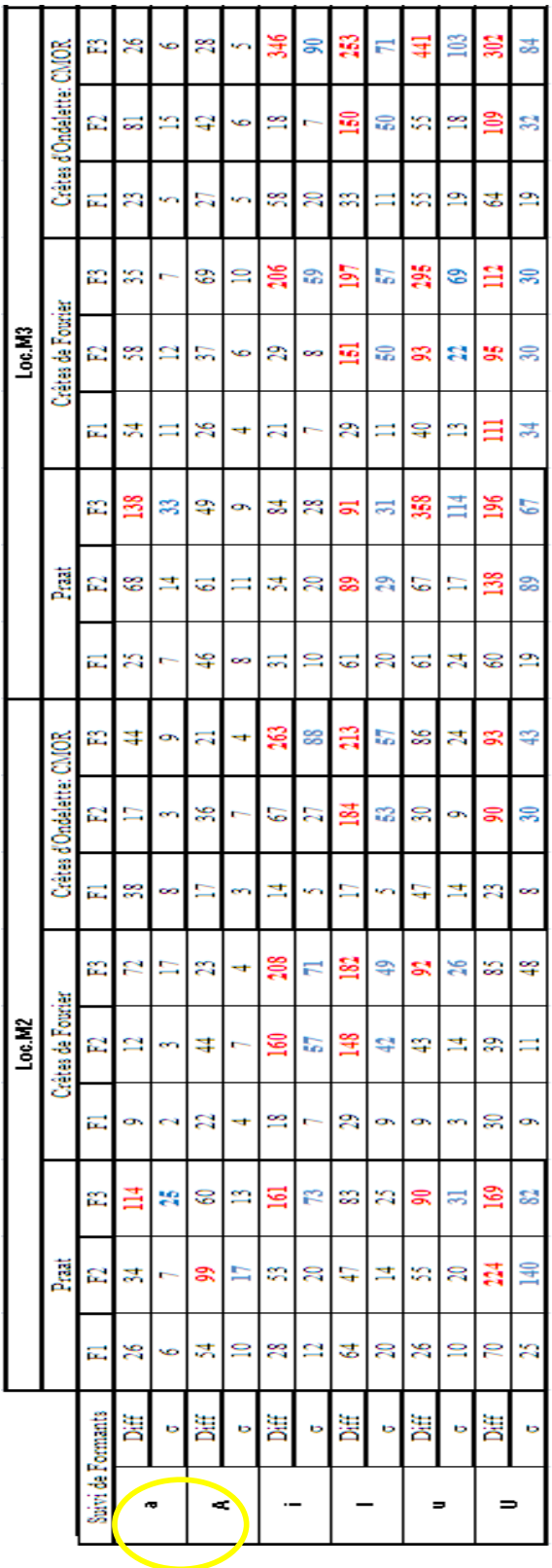


Figure 5. 7 : *Trajectoires formantiques estimées du signal sonore «*قَادَ الجَيْشَ.ا*» «**qa:da ljayŝa**» prononcée par la locutrice W4 (a) Etiquetage formantique manuel de référence,(b)Trajectoires formantiques estimées utilisant Praat (c)Trajectoires formantiques estimées utilisant les crêtes de Fourier et (d) Trajectoires formantiques estimées utilisant les crêtes d'ondelette cmor10-1*

Les tableaux 5.7 et 5.8 ci-dessous montrent les résultats obtenus sur les voyelles courtes /a, /i/ et /u/ et sur les voyelles longues /A/, /I/ et /U/ en calculant la moyenne de différence absolue (Eq.5.1) notée (Diff en Hz) et l'écart-type normalisé en fonction des valeurs de référence (Eq.5.2) notée (σ en %) pour chaque formant (F1, F2 et F3). Les occurrences des voyelles ont été extraites des quatre phrases présentées ci-dessous, par les locuteurs M2, M3 et M4.

- « عَرَفَ وَالِيًا وَقَائِدًا» « **ʕarafa wa:liyan wa qa:ʔidan** » (Il connait un gouverneur et un commandant)
- « هِيَ هُنَا لَقَدْ آبَتْ» « **Hiya huna: laqad ʒa:bat** ». (Elle était ici et elle était pieuse)
- « .أَسَرُونَا بِمُنْعَطَفٍ» « **ʒasaru:na: bimunʕatafin** » (Ils nous ont capturés dans un coin)
- « ؟أَتُؤْذِيهَا بِآلامِهِمْ» « ***ʒatuʒðiːha: biʒaːlaːmihim ?***» (Vous êtes en train de la blesser avec ses douleurs ?)



Suivi de Formants		Loc.M2									Loc.M3								
		Praat			Crêtes de Fourier			Crêtes d'Ondelettes: CWOR			Praat			Crêtes de Fourier			Crêtes d'Ondelettes: CWOR		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
a	Diff	26	34	114	9	12	72	38	17	44	25	68	138	54	58	35	23	81	26
	σ	6	7	25	2	3	17	8	3	9	7	14	33	11	12	7	5	15	6
A	Diff	54	99	60	22	44	23	17	36	21	46	61	49	26	37	69	27	42	28
	σ	10	17	13	4	7	4	3	7	4	8	11	9	4	6	10	5	6	5
i	Diff	28	53	161	18	160	208	14	67	263	31	54	84	21	29	206	58	18	346
	σ	12	20	73	7	57	71	5	27	88	10	20	28	7	8	89	20	7	90
I	Diff	64	47	83	29	148	182	17	184	213	61	89	91	29	151	197	33	150	253
	σ	20	14	25	9	42	49	5	53	67	20	29	31	11	50	67	11	50	71
u	Diff	26	55	90	9	43	92	47	30	86	61	67	358	40	93	295	55	55	441
	σ	10	20	31	3	14	26	14	9	24	24	17	114	13	22	69	19	18	103
U	Diff	70	124	169	30	39	85	23	90	93	60	138	196	111	95	112	64	109	302
	σ	25	140	82	9	11	48	8	30	43	19	89	67	34	30	30	19	32	84

Tableau 5. 7 : *Résultats obtenus sur les expressions des voyelles courtes et longues extraites des phrases de notre base de données prononcées par les locuteurs masculins M2 et M3*

		Loc.M4						Loc.M1					
		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR			Praat			Crêtes de Fourier			Crêtes d'Ondelette: MORLET		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
a	Diff	37	71	67	20	39	37	13	26	19	28	27	56	13	22	36	17	40	41
	σ	8	18	14	4	8	7	2	5	4	6	5	12	3	4	8	4	8	8
A	Diff	58	65	56	30	58	35	58	37	15	40	44	34	25	31	30	46	28	19
	σ	13	15	11	5	11	5	9	7	3	8	8	7	5	6	6	9	5	4
i	Diff	42	64	112	17	57	118	34	47	55	10	33	62	26	58	34	20	21	35
	σ	12	19	32	6	17	32	10	17	17	3	9	19	8	16	11	5	6	11
I	Diff	35	39	153	17	28	176	13	22	80	42	67	64	105	31	178	39	49	183
	σ	12	14	64	6	12	55	5	8	28	13	19	21	28	9	55	13	17	61
u	Diff	119	332	324	34	32	144	49	35	395	57	82	89	27	51	137	25	32	96
	σ	60	196	112	12	9	43	15	12	113	15	20	28	7	12	51	6	9	26
U	Diff	197	768	512	35	63	502	40	72	506	65	83	265	55	40	380	44	48	399
	σ	78	332	167	14	24	152	14	26	161	19	22	64	14	10	99	10	15	92

Tableau 5. 8 : *Résultats obtenus sur les expressions des voyelles courtes et longues extraites des phrases de notre base de données prononcées par les locuteurs masculin M4 et M1.*

➢ **Interprétations des résultats**

Les résultats présentés par les tableaux 5.7 et 5.8 sont les résultats obtenus sur les trois voyelles courtes /a/, /i/ et /u/ et les trois voyelles longues /A/, /I/ et /U/. Les occurrences de ces voyelles ont été extraites des phrases prononcées par trois locuteurs masculins M2, M3, M4 et M1. D'après ces résultats, on observe que le suivi est correct pour F1, F2 et F3 pour les voyelles /a/ et A/ (voir les cas encerclés en jaune dans les tableaux 5.7 et 5.8). Cette constatation est valable pour les trois méthodes et pour les trois locuteurs car il n'y a pratiquement pas d'erreurs (les valeurs en rouge) pour ces deux cas là. Par contre les résultats des voyelles /i/ et /u/ sont bons pour le suivi des formants F1 et F2 mais montrent qu'il y a des erreurs pour F3 et cela pour toutes les méthodes et tous les locuteurs. Par exemple dans le tableau 5.7 pour la voyelle /i/ et le locuteur M2, les trois erreurs sur F3 soit de 161, 208 et 263 Hz données respectivement pour les trois méthodes Praat, Fourier et ondelette. De même pour le cas de la voyelle /u/ prononcée par M3, les trois erreurs sur F3 sont de 358, 295 et 441 Hz données respectivement pour les trois méthodes Praat, Fourier et ondelette. A propos des voyelles longues /U/ et /I/, les erreurs concernent F2 et F3, et seul F1 est en général bien suivi comme cela apparaît dans le tableau 5.7. Par contre d'après les résultats du tableau 5.8 pour le locuteur M4, les erreurs sont concentrées sur F3 alors que le suivi de F1 et F2 est correct dans le cas des voyelles /i/ et /I/ pour les méthodes Praat et Fourier. En revanche la méthode par ondelette donne un bon suivi pour les trois formants.

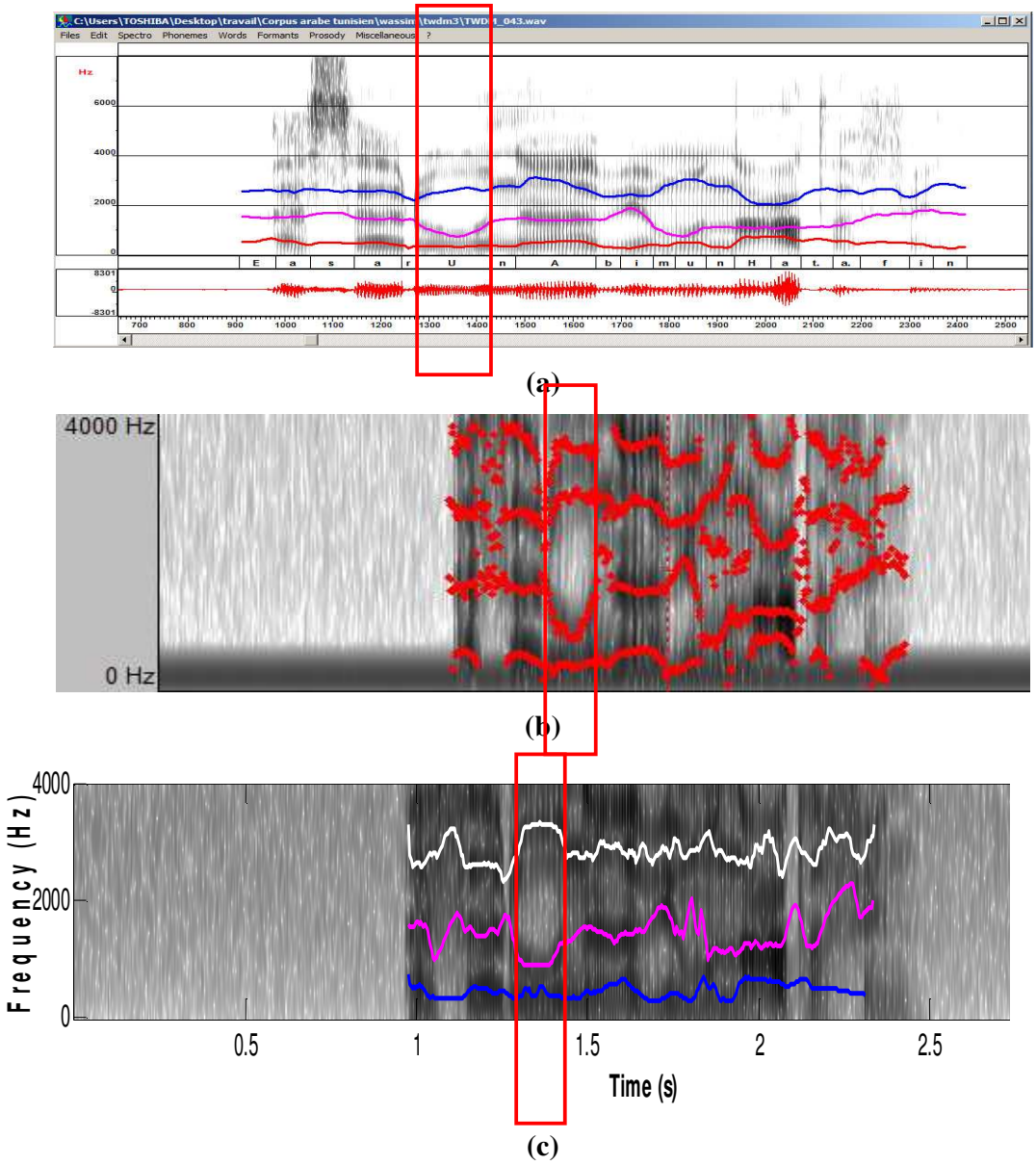
Dans le cas des voyelles /u/ et /U/ la méthode Praat présente un très mauvais suivi des formants F1, F2 et F3 (voir la zone encadrée en rouge dans le tableau 5.8 pour le locuteur M4) contrairement aux méthodes Fourier et ondelette qui présentent seulement quelques erreurs localisées sur F3.

D'après les résultats de suivi pour le locuteur M1, on remarque que le tableau 5.8 présente un suivi correct pour toutes les voyelles et par les trois méthodes, il y a seulement la présence de quelques erreurs concentrées sur F3 pour les voyelles /I/, /u/ et /U/ par les trois méthodes.

En conclusion on peut noter que d'après les résultats des deux tableaux 5.7 et 5.8, les erreurs se manifestent le plus souvent sur F3 pour les voyelles /i/ et /I/ et spécialement pour les voyelles /u/ et /U/. Cette observation vaut pour les trois méthodes quelque soit le locuteur et pour expliquer un peu ces erreurs, on a eu recours à une comparaison visuelle. D'après la figure 5.8, on remarque bien à travers les résultats de suivi des trois méthodes ainsi que la référence que l'énergie de formant F3 de la voyelle longue /U/ est très faible et à peine visible

ce qui explique le faux suivi (suivi de faux pics) que présente les trois méthodes à cause de l'absence des fréquences candidates correspondantes .

On constate aussi que les résultats des méthodes ondelette et Fourier ont été souvent proches dans certains cas.



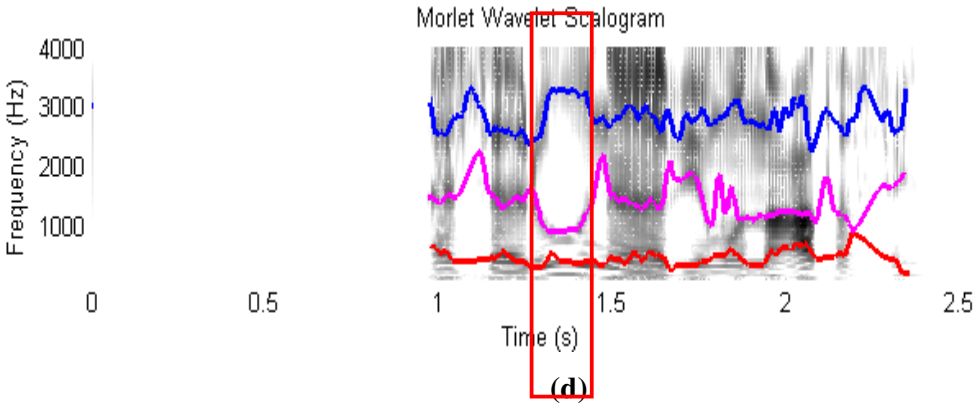
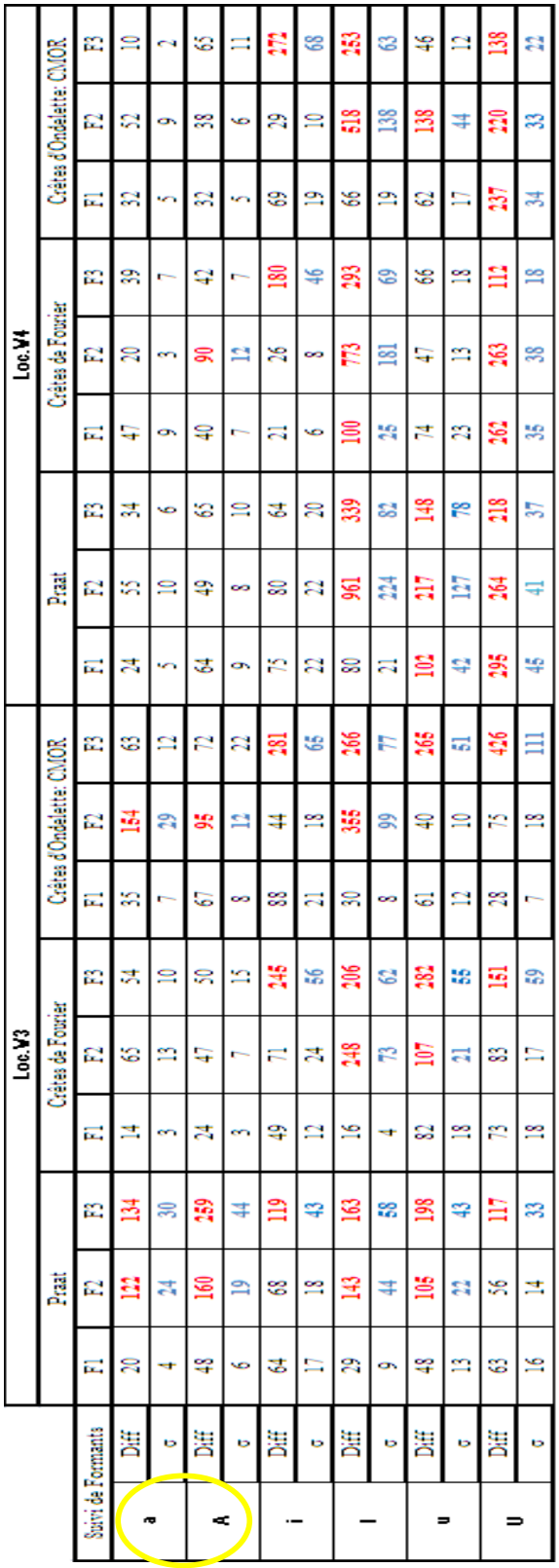


Figure 5. 8 : *Trajectoires formantiques estimées du signal sonore* « أسَرُوناَ بِمُنْعَطَفٍ. » « **ʒasaru:na: bimunʕatafin** » *prononcée par le locuteur M4 (a) Etiquetage formantique manuel de référence,(b)Trajectoires formantiques estimées utilisant Praat (c)Trajectoires formantiques estimées utilisant les crêtes de Fourier et (d) Trajectoires formantiques estimées utilisant les crêtes d'ondelette cmor10-1*

Les tableaux 5.9 et 5.10 ci-dessous montrent les résultats obtenus sur les voyelles courtes /a, /i/ et /u/ et sur les voyelles longues /A/, /I/ et /U/ en calculant la moyenne de différence absolue (Eq.5.1) notée (Diff en Hz) et l'écart-type normalisé en fonction des valeurs de référence (Eq.5.2) notée (σ en %) pour chaque formant (F1, F2 et F3). Les expressions des voyelles ont été extraites des mêmes phrases présentées précédemment dans le cas des locuteurs masculins mais prononcées par les locutrices W3, W4 et W5.



		Loc.W3									Loc.W4								
		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR			Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
a	Diff	20	122	134	14	65	54	35	154	63	24	55	34	47	20	39	32	52	10
	σ	4	24	30	3	13	10	7	29	12	5	10	6	9	3	7	5	9	2
A	Diff	48	160	259	24	47	50	67	95	72	64	49	65	40	90	42	32	38	65
	σ	6	19	44	3	7	15	8	12	22	9	8	10	7	12	7	5	6	11
i	Diff	64	68	119	49	71	245	88	44	281	75	80	64	21	26	180	69	29	272
	σ	17	18	43	12	24	56	21	18	65	22	22	20	6	8	46	19	10	68
I	Diff	29	143	163	16	248	206	30	365	266	80	961	339	100	773	293	66	818	253
	σ	9	44	58	4	73	62	8	99	77	21	224	82	25	181	69	19	138	63
u	Diff	48	105	198	82	107	282	61	40	265	102	217	148	74	47	66	62	138	46
	σ	13	22	43	18	21	55	12	10	51	42	127	78	23	13	18	17	44	12
U	Diff	63	56	117	73	83	161	28	75	426	295	264	218	262	263	112	237	220	138
	σ	16	14	33	18	17	59	7	18	111	45	41	37	35	38	18	34	33	22

Tableau 5. 9 : *Résultats obtenus sur les expressions des voyelles courtes et longues extraites des phrases de notre base de données prononcées par les locutrices W3 et W4.*

Loc.W5		Praat			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3
a	Diff	47	57	130	42	33	54	29	35	48
	σ	9	11	25	8	6	12	5	6	8
A	Diff	42	61	64	31	102	48	40	53	36
	σ	7	13	12	5	16	7	7	8	6
i	Diff	83	62	185	245	31	412	53	39	282
	σ	29	33	75	81	16	141	16	15	103
I	Diff	308	154	149	41	64	267	89	95	146
	σ	166	75	66	10	18	69	32	28	51
u	Diff	80	91	119	63	312	153	96	98	155
	σ	21	21	36	16	62	33	23	19	41
U	Diff	134	148	113	47	176	368	123	97	181
	σ	27	53	39	8	39	61	18	22	31

Tableau 5. 10 : *Résultats obtenus sur les expressions des voyelles courtes et longues extraites des phrases de notre base de données prononcées par la locutrice W5.*

➢ **Interprétations des résultats**

Les résultats présentés par les tableaux 5.9 et 5.10 sont les résultats obtenus sur les trois voyelles courtes /a/, /i/ et /u/ et les trois voyelles longues /A/, /I/ et /U/. Les expressions de ces voyelles ont été extraites des phrases prononcées par trois locutrices qui sont W3, W4 et W5.

Ces résultats montrent que le suivi de F1, F2 et F3 est correct pour les voyelles /a/ et A/ (voir les cas encerclés en jaune dans les tableaux 5.9 et 5.10). Cela vaut pour les trois

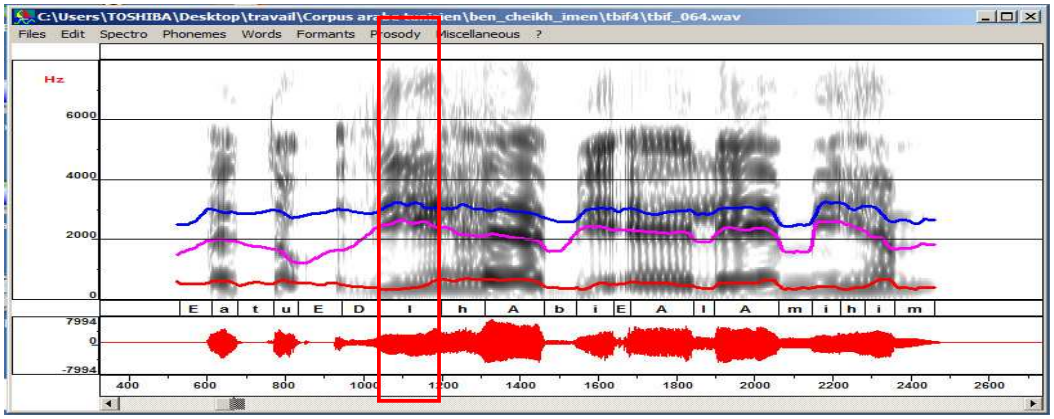
méthodes et les trois locutrices, à l'exception de la méthode LPC via Praat pour la locutrice W3 et les formants F2 et F3 (voir le tableau 5.9).

Dans le cas de la voyelle /i/, on note que le suivi est correct pour F1 et F2 et qu'il y a quelques erreurs sur F3 pour les trois méthodes et avec les trois locutrices.

Dans le cas de la voyelle /u/, les tableaux 5.9 et 5.10 montrent que le suivi de F1 et F2 est correct avec les deux méthodes crêtes de Fourier et crêtes d'ondelette mais qu'il y a souvent des erreurs sur F3. La méthode de Praat présente quand à elle des erreurs de suivi sur F2 et F3.

Dans le cas de la voyelle longue /I/, on remarque qu'on a des erreurs concernent les trois méthodes et essentiellement sur F2 et F3. En contrepartie le suivi de F1 est très bon. Pour expliquer un peu ces erreurs on eu recours à une comparaison visuelle. D'après les résultats présentée par la figure 5.9, on note que le suivi des formants F2 et F3 de la voyelle longue /I/ fait par les trois méthodes n'est pas lisse et ne suit pas correctement l'énergie comparant au suivi référence ce qui confirme les erreurs concentrées sur F3 et surtout sur F2 pour la locutrice W3 (voir tableau 5.9) et pour les autres locutrices. Ces erreurs peuvent être expliquées par le fait que cette voyelle est suivi par la consonne fricative non voisée /h/ qui au sein de la phrase se comporte comme si elle est voisée et elle se caractérise par des pseudo-formants qui se forment de la voyelle qui la précède ce qui engendre de faux formants et cela influe sur le suivi de formants de la voyelle qui la précède qui est dans notre cas la voyelle /I/ dû à la présence de faux pics ce qu'on voit bien dans la cas de la méthode LPC faite par le logiciel Praat.

Dans le cas de la voyelle longue /U/, on note que les erreurs portent sur les trois formants, les trois méthodes et les trois locutrices.



(a)

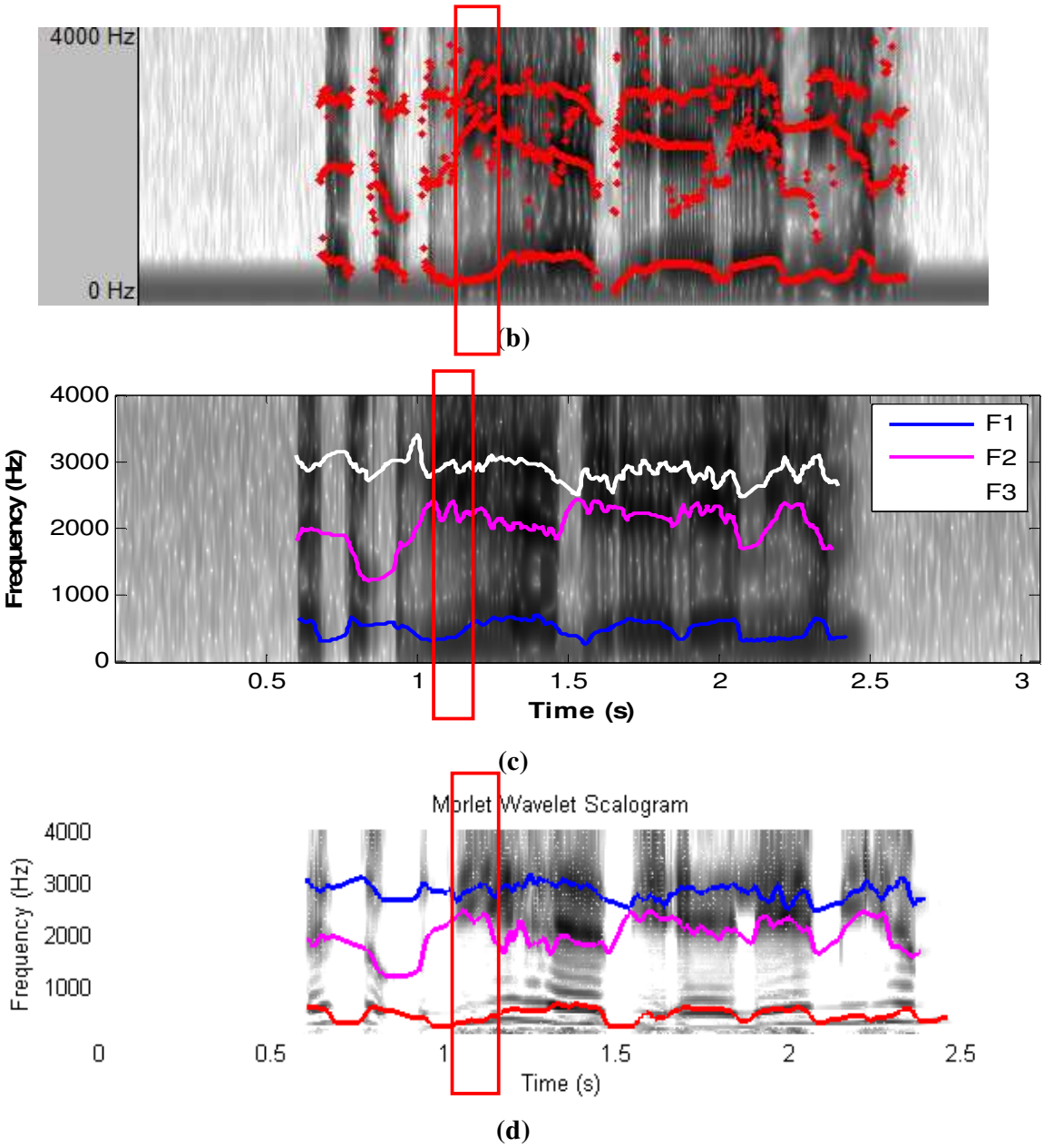


Figure 5. 9 : *Trajectoires formantiques estimées du signal sonore «* آتُوْذِيهَا بِآلامِهِمْ؟ *» « **ʒatuʒðiːhaː biʒaːlaːmihim ?** » prononcée par la locutrice W3 (a) Etiquetage formantique manuel de référence,(b)Trajectoires formantiques estimées utilisant Praat (c)Trajectoires formantiques estimées utilisant les crêtes de Fourier et (d) Trajectoires formantiques estimées utilisant les crêtes d'ondelette cmor10-1*

5.4.2. Etude et évaluations quantitatives sur différentes phrases

Dans cette section, nous présentons tout d'abord, la méthode de l'analyse LPC combinée à des bancs de filtres proposée par (Mustafa.K 2006) et utilisée ici comme méthode de comparaison. Ensuite, nous discutons les résultats obtenus sur différentes phrases en

interprétant les valeurs de la différence moyenne et de l'écart type calculées pour chaque méthode automatique de suivi de formants.

5.4.2.1. Méthode de suivi de formants proposée par Mustafa Kamran

La méthode de Mustafa Kamran est basée sur le filtrage de la parole à partir d'un banc de filtres. Le but de ce filtrage est de limiter la région spectrale de l'estimation pour chaque formant pour éviter la fusion des trajectoires formantiques et l'influence du bruit sur l'estimation. La figure présentée ci-dessous montre les différentes étapes de cette technique. (voir Fig.5.10)

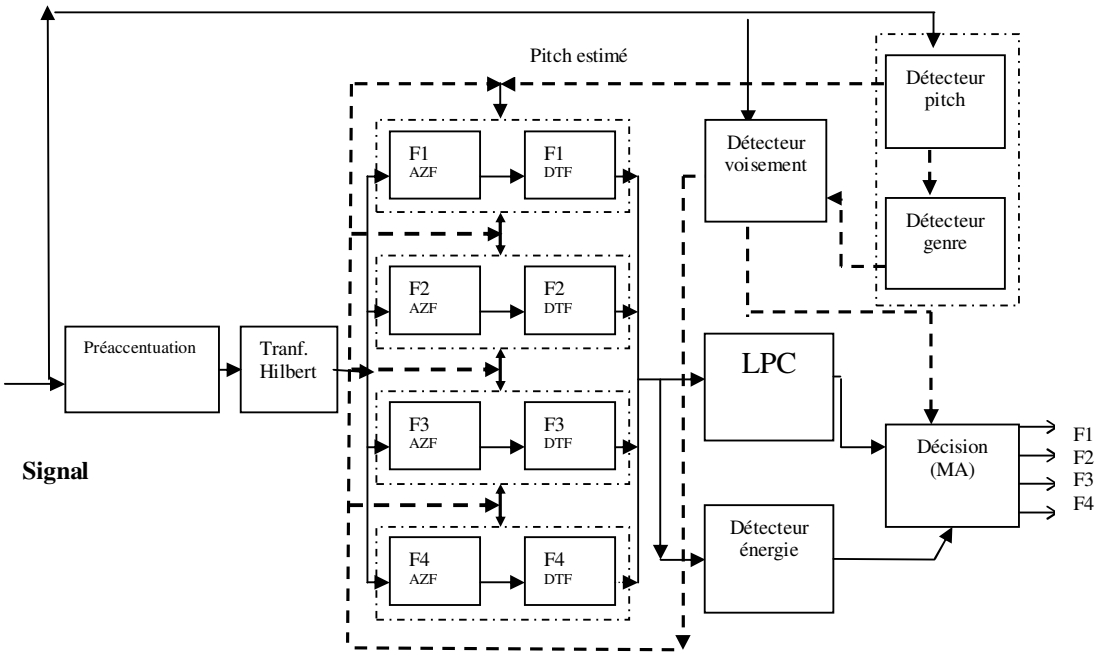


Figure 5. 10 : *Diagramme de l'approche proposée par* (Mustafa.K 2006)

D'abord le signal est préaccentué en utilisant des filtres passe haut pour accentuer les fréquences de faible énergie. Ensuite, la transformée de Hilbert est appliquée sur le signal pour augmenter la précision spectrale.

Le système est constitué d'un banc de filtres pour le préfiltrage de chaque formant (F1, F2, F3 et F4). Ce banc est formé d'un filtre des zéros (AZF) présents dans le signal de parole et d'un filtre de suivi dynamique (DTF) pour garantir la continuité du suivi au cours de temps en fournissant à chaque fois le pôle de la dernière estimation déjà faite.

La sortie de chaque banc de filtre est une bande passante pour chaque formant afin de supprimer les effets des formants voisins (à travers AZF) et pour que l'estimation des formants soit plus précise et plus robuste au bruit blanc gaussien.

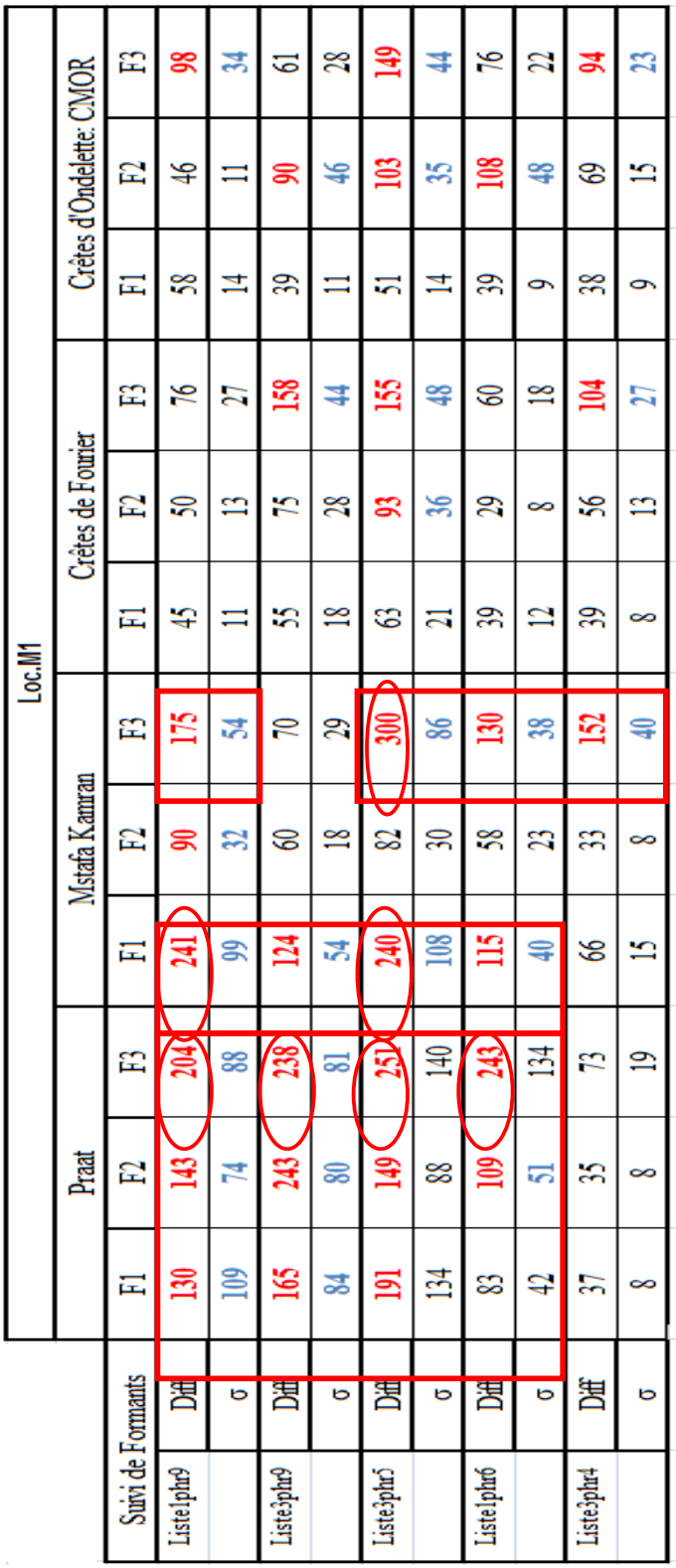
Les coefficients LPC sont ainsi calculés pour les sons voisés à partir du signal analytique du filtre de chaque formant pour estimer les fréquences candidates.

Puisque les quatre formants varient avec le temps, les filtres sont adaptés pour suivre une bande de fréquence en changeant la localisation de leurs pôles et zéros.

Cette technique est améliorée en utilisant un détecteur de voisement pour trouver les sons voisés ou non voisés. Lorsque le signal est non voisé ou lorsque l'énergie dans le signal est insuffisante, les fréquences des formants estimées prennent la valeur moyenne de leur déplacements (Moving Average) ce qui assure la continuité de suivi avec un minimum d'erreur après des segments de parole non voisés ou bien de basse énergie. Le détecteur de voisement s'adapte en fonction du sexe du locuteur ce qui améliore la détection du voisement. Les seuils utilisés des bandes d'énergie sont aussi adaptatifs afin qu'ils puissent s'adapter aux changements à long terme des niveaux d'énergie de chaque région de formant variable au cours de temps.

5.4.2.2. Interprétations et discussions des résultats

Les tableaux 5.11, 5.12 et 5.13 ci-dessous présentent les résultats des tests de notre nouvel algorithme crêtes d'ondelette sur la parole continue. Nous avons comparé les résultats de suivi obtenus avec les trois méthodes crêtes de Fourier, la méthode de Mustafa Kamran basée sur l'analyse LPC et l'analyse LPC dans le logiciel Praat. Les tests ont été faits sur des phrases voisées de notre base de données en langue arabe standard (voir Annexe B). Ces phrases ont été choisies au hasard, segmentées selon leurs parties voisées tout en éliminant les zones de silences et prononcées par deux locuteurs masculins (M1 et M4) et une locutrice W3. Le choix des locuteurs a été pris au hasard, faute de temps nous n'avons pas pu faire les tests sur tous les locuteurs et les locutrices.



Loc.M1		Praat			Mstafa Kamran			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Liste1phr9	Diff	130	143	204	241	90	175	45	50	76	58	46	98
	σ	109	74	88	99	32	54	11	13	27	14	11	34
Liste2phr9	Diff	165	243	238	124	60	70	55	75	158	39	90	61
	σ	84	80	81	54	18	29	18	28	44	11	46	28
Liste2phr5	Diff	191	149	251	240	82	300	63	93	155	51	103	149
	σ	134	88	140	108	30	86	21	36	48	14	35	44
Liste1phr6	Diff	83	109	243	115	58	130	39	29	60	39	108	76
	σ	42	51	134	40	23	38	12	8	18	9	48	22
Liste2phr4	Diff	37	35	73	66	33	152	39	56	104	38	69	94
	σ	8	8	19	15	8	40	8	13	27	9	15	23

Tableau 5. 11 : *Résultats obtenus sur cinq phrases voisées de notre base de données prononcées par le locuteur masculin M1.*

Loc.M4		Praat			Mstafa Kamran			Crêtes de Fourier			Crêtes d'Ondelette: CMOR		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Listeóphr2	Diff	69	768	141	212	116	143	74	61	106	60	63	123
	σ	66	606	74	119	58	56	33	24	41	19	28	55
Listeóphr4	Diff	71	192	153	300	97	497	47	73	163	51	88	142
	σ	59	256	56	156	34	106	16	27	47	15	27	43
Listeóphr9	Diff	44	480	165	62	80	303	61	43	163	55	39	151
	σ	12	210	51	27	38	70	18	13	46	13	10	58
Liste7phr2	Diff	44	155	123	104	69	148	41	60	162	65	76	114
	σ	26	117	48	50	30	51	17	28	56	19	37	61
Lsite7phr3	Diff	138	89	120	190	80	140	37	50	76	71	46	80
	σ	121	65	82	126	40	66	9	16	30	17	14	34

Tableau 5. 12 : *Résultats obtenus sur cinq phrases voisées de notre base de données prononcées par le locuteur masculin M4.*

Loc.W3												
		Praat			Mistafa Kamran			Crêtes de Fourier			Crêtes d'Ondelette: CMOR	
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
Liste4phr9	Diff	68	46	157	76	56	60	68	63	65	68	90	62
	σ	38	10	64	29	15	16	24	12	13	16	20	12
Liste5phr5	Diff	56	55	108	311	282	117	85	70	105	62	178	100
	σ	23	13	46	82	67	34	25	19	26	18	36	26
Liste5phr10	Diff	34	69	102	110	85	79	91	147	74	105	186	72
	σ	7	23	27	28	20	12	15	27	12	15	31	14
Liste2phr4	Diff	32	52	185	151	543	247	55	127	93	77	173	87
	σ	7	27	57	46	165	88	17	41	25	17	45	25
Liste1phr10	Diff	51	118	268	131	106	52	109	151	110	60	102	97
	σ	15	42	103	47	36	13	35	53	25	16	30	22

Tableau 5. 13 : *Résultats obtenus sur cinq phrases voisées de notre base de données prononcées par la locutrice W3.*

- **Interprétations des résultats**

Le tableau 5.11, montre que le suivi est globalement pertinent pour F1, F2 et F3 pratiquement pour les cinq phrases prononcées par le locuteur M1 et cela pour les deux méthodes crêtes de Fourier et crêtes d'ondelette. Il y a seulement quelques petites erreurs de suivi sur F2 ou F3 par la méthode crêtes d'ondelette et quelques erreurs de suivi sur F3 par la méthode crêtes de Fourier. En revanche, la méthode LPC de Mustafa Kamran donne un bon suivi pour F2 car il y a plusieurs erreurs sur F1 et F3 pratiquement pour les cinq phrases (voir les cases encadrés en rouge dans le tableau 5.11). En contrepartie les résultats de la méthode LPC exécutés par le logiciel Praat présentent un mauvais suivi pour les trois formants sur les quatre premières phrases par rapport aux autres méthodes, il y a même plusieurs erreurs qui dépassent les 200Hz sur F3.

D'après les résultats présentés dans le tableau 5.12, on note que le suivi des formants F1 et F2 est correct pour les quatre premières phrases prononcées par le locuteur M4 et on a vérifié cela pour les deux méthodes crêtes de Fourier et crêtes d'ondelette. En revanche on voit des erreurs sur F3 pour les quatre premières phrases (voir les cases encadrés en rouge dans le tableau 5.12) et que la cinquième phrase (Liste7phr3) donne un bon suivi pour ces deux méthodes.

Comme on le voit dans ce qui précède, les deux méthodes crêtes de Fourier et crêtes d'ondelette donnent de meilleurs résultats que celle de Mustafa Kamran et celle de LPC dans le logiciel Praat (voir les erreurs encadrées en rouge dans le tableau 5.12). Cette dernière présente un suivi correct sur F1 et des erreurs concentrées sur F2 et F3 pour les cinq phrases.

Les résultats obtenus pour la locutrice W3 confirment la tendance observée sur les locuteurs. En effet, les méthodes de suivi par crêtes de Fourier ou d'ondelettes donnent de meilleurs résultats que celle de Mustafa Kamran, même si globalement les résultats sont moins bons que pour les locuteurs masculins. En revanche dans ce cas là, le suivi des formants F1 et F2 de la méthode LPC de Praat se montre meilleur que ceux des autres méthodes sauf que cette méthode présente des erreurs concentrées sur F3 (voir le tableau 5.13).

5.5. Etude et évaluations quantitatives de l'algorithme de suivi de formants utilisant la programmation dynamique combiné avec le filtrage de Kalman

Afin d'évaluer notre nouvel algorithme de suivi de formants crêtes de Fourier utilisant la programmation dynamique combiné avec le filtrage de Kalman, nous avons fait des tests sur les mêmes phrases utilisées dans la section précédente prononcées par les mêmes locuteurs en calculant la moyenne de différence absolue (Eq.5.1) notée (Diff en Hz) et l'écart-type normalisé en fonction des valeurs de référence (Eq.5.2) notée (σ en %) pour chaque formant (F1, F2 et F3). A ce stade les résultats quantitatifs de suivi de nouvel algorithme ont été comparés à ceux des trois méthodes crêtes de Fourier utilisant le calcul de centre de gravité, l'analyse LPC combinée à des bancs de filtres de Mustafa Kamran (Mustafa.K 2006) et la méthode LPC dans le logiciel Praat.

Les tableaux 5.14, 5.15 et 5.16 ci-dessous présentent les résultats quantitatifs des tests de notre nouvel algorithme crêtes de Fourier utilisant la programmation dynamique combiné avec le filtrage de Kalman sur la parole continue ainsi que les résultats des deux autres méthodes de comparaison. En fait, on a gardé les mêmes résultats trouvés précédemment pour ces deux méthodes pour pouvoir les comparer avec le nouvel algorithme.

Loc.M1												
		Mstafa Kamran			Fourier avec centre de Gravité			Fourier avec filtre de Kalman				
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3		
Liste1phr9	Diff	241	90	175	45	50	76	39	75	110		
	σ	99	32	54	11	13	27	11	22	40		
Liste3phr9	Diff	124	60	70	55	75	158	47	50	105		
	σ	54	18	29	18	28	44	16	22	36		
Liste3phr5	Diff	240	82	300	63	93	155	78	100	290		
	σ	108	30	86	21	36	48	25	41	86		
Liste1phr6	Diff	115	58	130	39	29	60	54	70	93		
	σ	40	23	38	12	8	18	16	27	30		
Liste3phr4	Diff	66	33	152	39	56	104	42	82	110		
	σ	15	8	40	8	13	27	9	29	28		

Tableau 5. 14 : *Résultats obtenus sur cinq phrases sonores de notre base de donnée prononcées par le locuteur masculin M1.*

Loc.M4											
	Mstafa Kamran			Fourier avec centre de Gravité			Fourier avec filtre de Kalman				
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3	
Liste6phr2	Diff	212	116	143	74	61	106	80	167	185	
	σ	119	58	56	33	24	41	34	89	77	
Liste6phr4	Diff	300	97	497	47	73	163	65	86	455	
	σ	156	34	106	16	27	47	18	24	90	
Liste6phr9	Diff	62	80	303	61	43	163	99	72	113	
	σ	27	38	70	18	13	46	28	31	32	
Liste7phr2	Diff	104	69	148	41	60	162	42	96	195	
	σ	50	30	51	17	28	56	19	54	67	
Lsite7phr3	Diff	190	80	140	37	50	76	40	45	59	
	σ	126	40	66	9	16	30	10	9	26	

Tableau 5. 15 : *Résultats obtenus sur cinq phrases sonores de notre base de donnée prononcées par le locuteur masculin M4.*

Loc.W3											
		Mstafa Kamran			Fourier avec centre de Gravité			Fourier avec filtre de Kalman		
Suivi de Formants		F1	F2	F3	F1	F2	F3	F1	F2	F3
Liste4phr9	Diff	76	56	60	68	63	65	78	62	51
	σ	29	15	16	24	12	13	21	15	10
Liste5phr5	Diff	311	282	117	85	70	105	89	71	108
	σ	82	67	34	25	19	26	32	19	26
Liste5phr10	Diff	110	85	79	91	147	74	89	128	87
	σ	28	20	12	15	27	12	16	26	15
Liste2phr4	Diff	151	543	247	55	127	93	45	79	97
	σ	46	165	88	17	41	25	14	23	31
Liste1phr10	Diff	131	106	52	109	151	110	120	128	106
	σ	47	36	13	35	53	25	39	44	24

Tableau 5. 16 : *Résultats obtenus sur cinq phrases sonores de notre base de donnée prononcées par la locutrice W3.*

- **Interprétations des résultats**

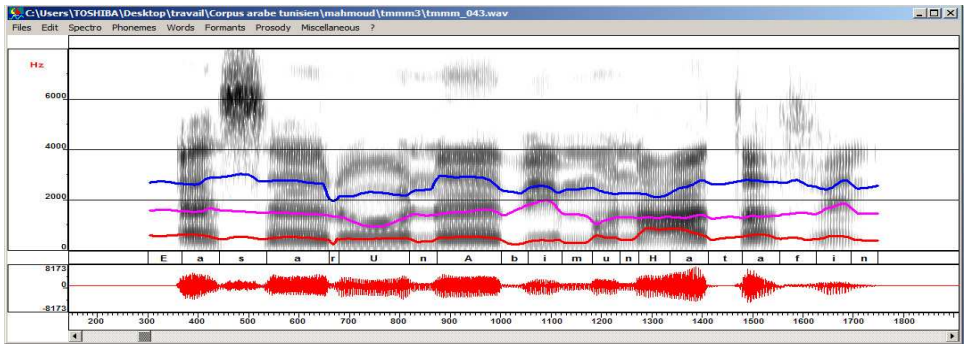
D'après les résultats présentés dans le tableau 5.14, on note bien qu'on a un suivi pertinent des formants F1 et F2 pratiquement pour les cinq phrases prononcées par le locuteur M1 et on a vérifié cela pour l'algorithme de suivi crêtes de Fourier utilisant le calcul de centre de gravité et celui des crêtes de Fourier utilisant le filtrage de Kalman. On remarque de même qu'il y a certains résultats de ces deux méthodes qui sont parfois très proches (voir les cases encadrées en jaune dans le tableau 5.14), par exemple, dans le cas de la phrase liste1phr9 en F1, les valeurs respectives de différence moyenne absolue et de l'écart type sont comme suit (Diff=45Hz et σ=11%) pour la méthode Fourier avec centre de gravité et les valeurs statistiques de la méthode Fourier avec filtrage de Kalman sont comme suit (Diff=39Hz et σ=11%). En revanche, il y a quelques erreurs de suivi sur F3 par la méthode crêtes de Fourier et plus d'erreurs par le nouvel algorithme utilisant le filtrage de Kalman. Par contre, comme on a déjà mentionné précédemment que la méthode LPC de Mustafa Kamran donne un bon suivi seulement sur F2 pratiquement sur les cinq phrases, cependant, on note bien que les résultats du nouvel algorithme utilisant le filtrage de Kalman sont très proches à ceux de la méthode Fourier avec centre de gravité et ils sont notamment meilleurs que ceux de la méthode LPC de Mustafa Kamran et ceux de la méthode LPC de Praat (voir le tableau 5.11). Ces résultats sont bien confirmés en comparant visuellement les résultats des différentes méthodes (voir l'exemple présenté par la figure 5.11).

D'après les résultats présentés dans le tableau 5.15, on note que le suivi des formants F1, F2 est correct pour les quatre premières phrases prononcées par le locuteur M4 et on a vérifié cela pour les deux méthodes Fourier avec centre de gravité et Fourier avec filtrage de Kalman. En revanche on voit des erreurs sur F3 pour les quatre premières phrases (voir les cases encadrés en rouge dans le tableau 5.15) et que la cinquième phrase (Liste7phr3) donne un bon suivi pour ces deux méthodes.

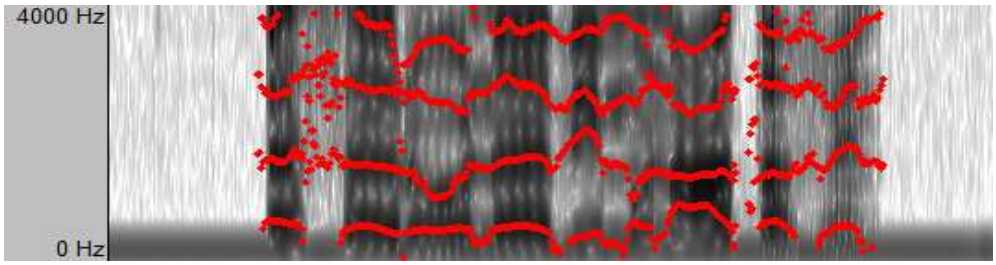
Comme on le voit dans ce qui précède, les deux méthodes Fourier avec centre de gravité et Fourier avec filtrage de Kalman donnent de meilleurs résultats que celle de Mustafa Kamran et celle de la méthode LPC dans le logiciel Praat (voir le tableau 5.12)

Les résultats obtenus dans le tableau 5.16 pour la locutrice W3 confirment la tendance observée sur les locuteurs. En effet, même remarque que tout à l'heure, les deux méthodes de suivi par crêtes de Fourier donnent de meilleurs résultats que celle de Mustafa Kamran qui présente des erreurs très élevées qui dépasse les 200 Hz et cela est bien confirmé en comparant visuellement les résultats des différentes méthodes (voir l'exemple présenté par la

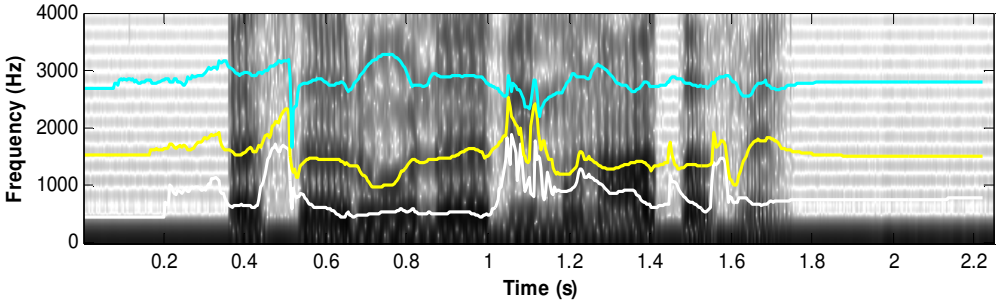
figure 5.12), même si globalement les résultats sont moins bons que pour les locuteurs masculins.



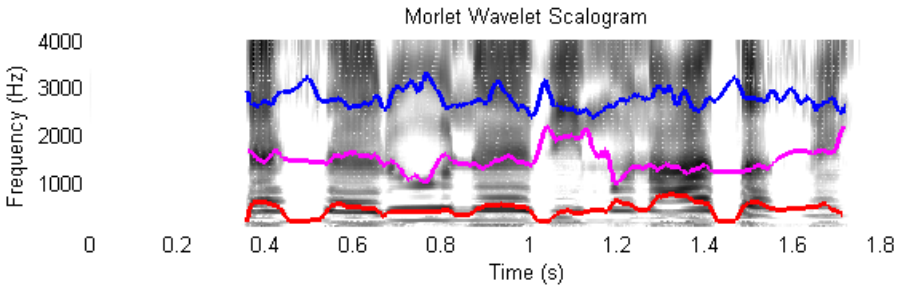
(a)



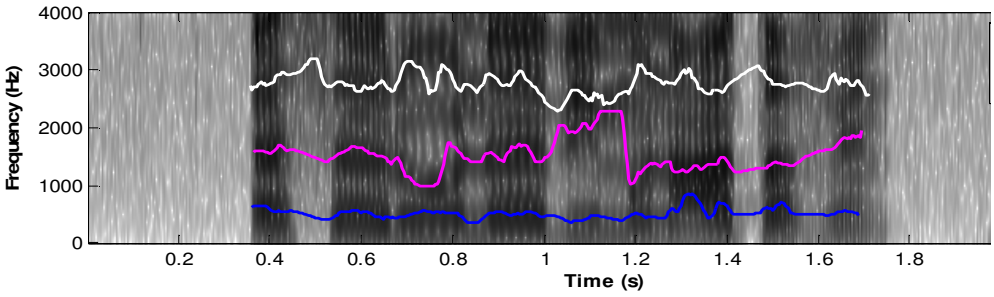
(b)



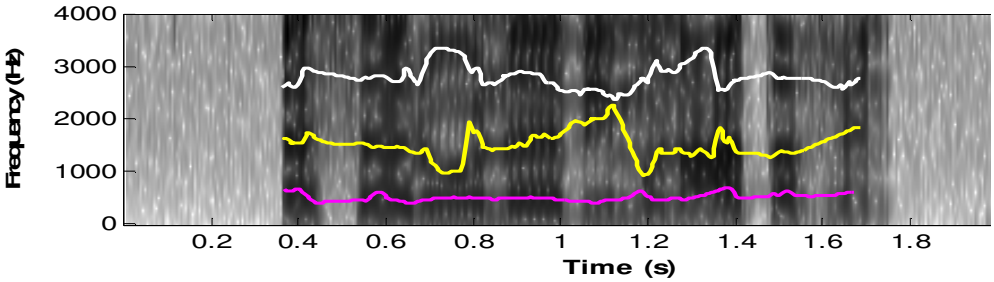
(c)



(d)

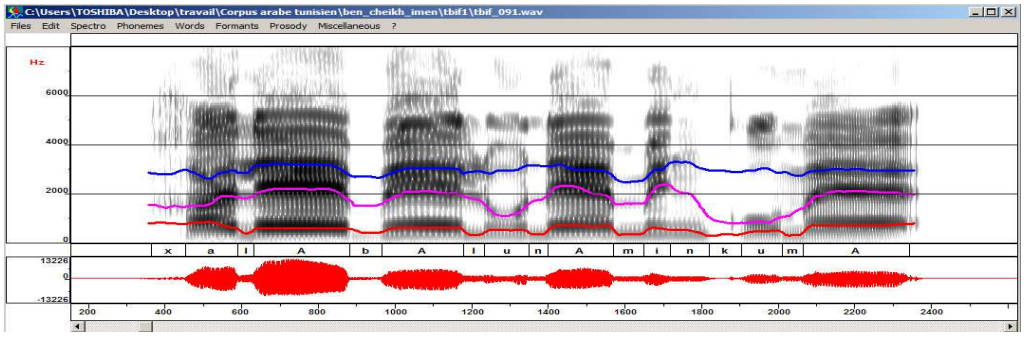


(e)

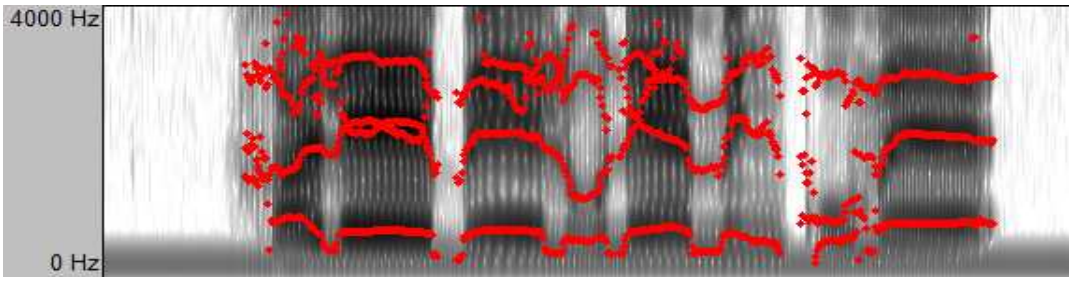


(f)

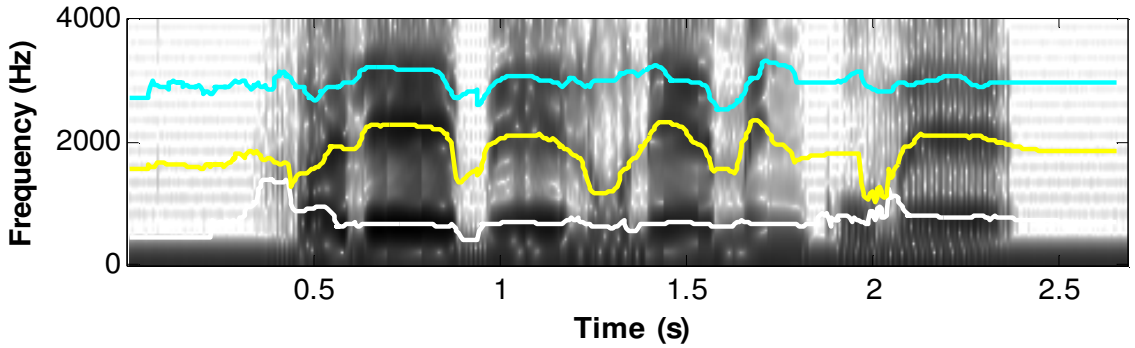
Figure 5. 11 : *Trajectoires formantiques estimées du signal sonore (liste3phr5)* « أسَرُونَا بِمُنْعَطَفٍ. » « **ʒasaru:na: bimunɛatafin** » *prononcée par le locuteur M1 (a) Etiquetage formantique manuel de référence,(b)Trajectoires formantiques estimées utilisant Praat, (c) Trajectoires formantiques estimées utilisant la méthode de Mustafa Kamran (d) Trajectoires formantiques estimées utilisant les crêtes d'ondelette cmor10-1 (e)Trajectoires formantiques estimées utilisant les crêtes de Fourier avec centre de gravité (f)Trajectoires formantiques estimées utilisant les crêtes de Fourier avec filtrage de Kalman*



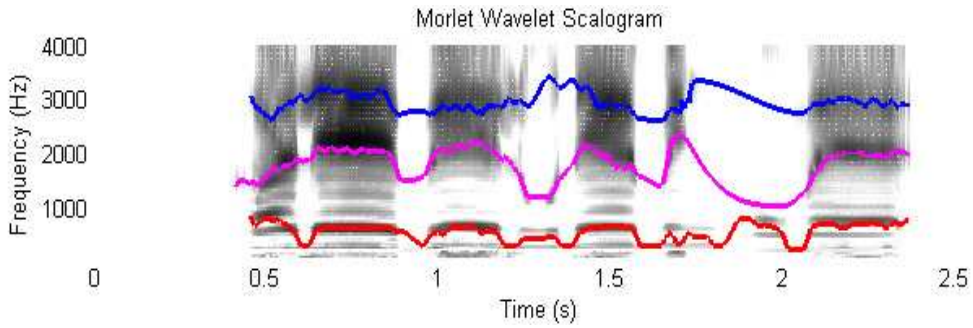
(a)



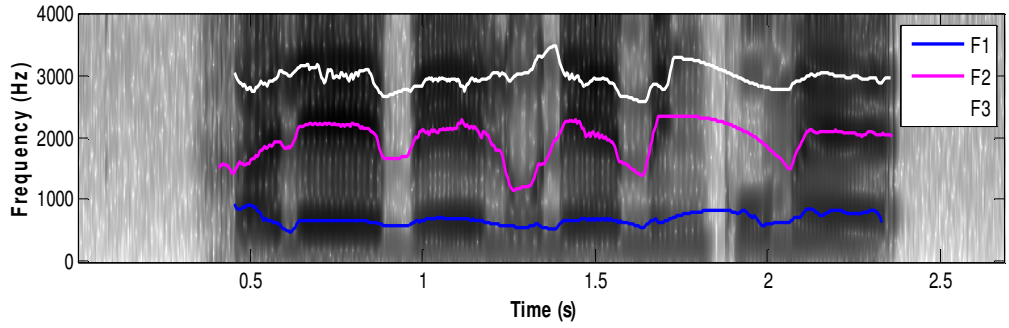
(b)



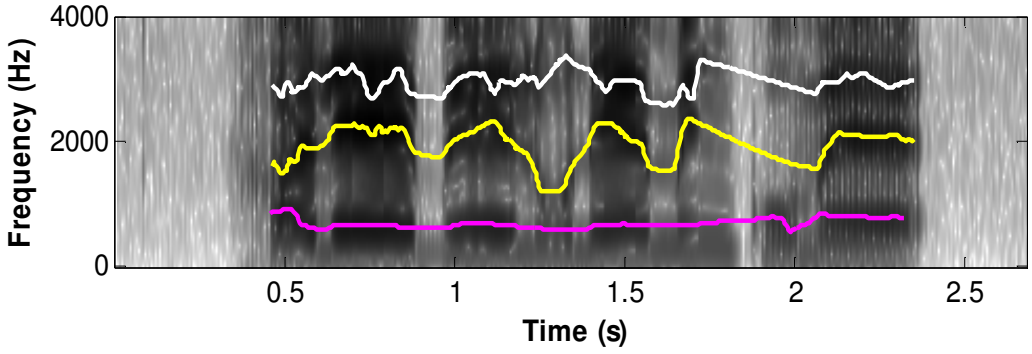
(c)



(d)



(e)



(f)

Figure 5. 12 : *Trajectoires formantiques estimées du signal sonore (liste1phr10) «* خَلاَ بَالُنَا مِنْكُمَا*» « **χala: ba:luna: minkuma:**» prononcée par le locutrice W3 (a) Etiquetage*

formantique manuel de référence,(b)Trajectoires formantiques estimées utilisant Praat, (c) Trajectoires formantiques estimées utilisant la méthode de Mustafa Kamran (d) Trajectoires formantiques estimées utilisant les crêtes d'ondelette cmor10-1 (e)Trajectoires formantiques estimées utilisant les crêtes de Fourier avec centre de gravité (f)Trajectoires formantiques estimées utilisant les crêtes de Fourier avec filtrage de Kalman

5.5. Conclusion

Nous avons présenté dans ce chapitre les résultats et les interprétations en testant nos deux nouvelles approches de suivi de formants. La première est basée sur la détection des crêtes d'ondelette en calculant le centre de gravité de l'ensemble des fréquences candidates comme contrainte de suivi et la deuxième se base sur la détection des crêtes de Fourier et la classification des fréquences candidates de chaque formants en utilisant la programmation dynamique en combinaison avec le filtrage de Kalman pour le suivi des trajectoires. La première approche est testée sur des signaux synthétiques ensuite sur des signaux réels de notre corpus étiqueté en testant trois types d'ondelettes.

Nous avons fait après, une évaluation quantitative des deux approches séparément en les comparant avec d'autres méthodes automatiques de suivi de formants avec différents locuteurs (masculins et féminins) et en prenant les signaux étiquetés issus de la base élaborée comme référence tout en mettant l'utilité de notre base de données étiquetées manuellement.

Dans le cas des locuteurs masculins, les résultats des deux nouvelles approches sont notamment meilleurs que ceux de la méthode LPC de Mustafa Kamran et ceux de Praat même si elles présentent souvent quelques erreurs sur F3.

Les résultats obtenus dans le cas des locutrices féminins confirment la tendance observée sur les locuteurs. En effet les deux nouvelles approches de suivi donnent de meilleurs résultats que celle de Mustafa Kamran qui présente des erreurs très élevées qui dépasse les 200 Hz, même si globalement les résultats sont moins bons que pour les locuteurs masculins.

Conclusion et Perspectives

L'objectif de ce travail est l'amélioration des performances des algorithmes de suivi des formants.

Afin de tenir compte des finalités poursuivies, nous avons entamé ce travail par l'analyse des différentes techniques classiques utilisées dans le suivi automatique des formants en plus d'autres méthodes plus récentes. Cette analyse nous a permis de constater que l'estimation automatique des formants reste délicate malgré l'emploi de diverses techniques complexes. Parmi ces méthodes nous pouvons citer : les méthodes paramétriques de l'extraction des formants telles que les techniques basées sur la prédiction linéaire, les vecteurs MFCC, l'analyse cepstrale et les techniques basées sur les modèles auditifs. Nous avons exploré aussi les méthodes traditionnelles de suivi des formants qui sont basées sur les algorithmes de la programmation dynamique, celles basées sur le filtrage de Kalman et d'autres basées sur les modèles de Markov cachés. Les modèles d'appariement des trajectoires de suivi de formants a connu des évolutions marquantes depuis les modèles HMM aux modèles dynamiques cachés HDM (Richards.H.B 1999) pour décrire la structure dynamique du son. Pour assurer la continuité du suivi et pour faire une bonne estimation des VTR, Li Deng a suggéré l'utilisation des modèles dynamiques cachés des VTR pour compenser la probabilité d'avoir des fréquences candidates manquantes pendant le suivi.

Nous avons étudié aussi d'autres méthodes basées sur l'inclusion de l'information du contexte phonémique dans l'estimation et le suivi de formants (Lee.M 2005) ; c'est une tâche qui repose sur l'alignement forcé et la segmentation phonémique.

Malgré l'importance du rôle joué par les fréquences des formants pour la perception et le traitement de la parole, on constate que les bases de données de référence sont très rares surtout en langue arabe. Comme ces bases sont nécessaires à l'évaluation quantitative des techniques automatiques de suivi des formants, nous nous sommes intéressés dans la suite de ce travail à la préparation de notre base de données en enregistrant un corpus phonétiquement équilibré en langue arabe et en élaborant un étiquetage manuel phonétique et formantique des différents signaux de cette base. A ce stade de travail, nous avons rencontré plusieurs difficultés spécialement pour les consonnes voisées. Nous avons donc fait appel à des experts en phonétique. Cette base de données est destinée à être utilisée par la suite comme base de référence dans le but d'évaluer les algorithmes automatiques de suivi de formants testés sur des signaux en langue arabe.

Suite à ce travail, nous avons présenté notre contribution qui est la méthode de suivi de formants par crêtes de Fourier (maxima de spectrogramme) en utilisant comme contrainte de suivi le calcul de centre de gravité de la combinaison des fréquences candidates pour chaque formant. Vu les limites de la transformée de Fourier fenêtrée, nous avons étendu nos recherches à la transformée en ondelette c'est-à-dire s'étendre de l'analyse monorésolution à l'analyse multirésolution. Nous avons implémenté alors une première approche basée sur la détection des crêtes d'ondelette qui sont les maxima du scalogramme en testant trois types d'ondelettes complexes qui sont : Morlet Complexe, Shanon et Frequency B-Spline. Pour réaliser le suivi des trajectoires fréquentielles, nous avons utilisé le calcul du centre de gravité de la combinaison des fréquences formantiques candidates comme première approche de suivi. Ensuite, nous avons implémenté une deuxième approche de suivi basée sur la programmation dynamique combiné avec le filtrage de Kalman.

Finalement et dans une dernière partie de ce travail, nous avons fait une étude exploratoire en utilisant notre corpus étiqueté manuellement comme référence pour évaluer quantitativement nos deux nouvelles approches par rapport à d'autres méthodes automatiques de suivi de formants.

Nous avons testé la première approche par détection des crêtes ondelette, utilisant le calcul de centre de gravité, sur des signaux synthétiques ensuite sur des signaux réels de notre corpus étiqueté en testant trois types d'ondelettes complexes qui sont : cmor, fbsp et shan.

Suite à ces différents tests, nous avons constaté que les résultats de ces différents tests de la méthode basée sur les ondelettes cmor et fbsp sont globalement presque similaires. Par contre lorsqu'on calcule le scalogramme en utilisant l'ondelette shan, nous constatons que le compromis entre la résolution du scalogramme et l'estimation des fréquences des formants est moins pertinent. On a remarqué que les formants ne sont pas bien mis en évidence et qu'il y a une apparition des harmoniques.

Il apparaît donc que le suivi de formants et la résolution des scalogrammes donnés par les ondelettes cmor et fbsp sont meilleurs qu'avec l'ondelette shan.

Afin d'évaluer quantitativement notre nouvel algorithme de suivi de formants par détection des crêtes d'ondelette en utilisant le calcul de centre de gravité, nous avons fait plusieurs tests avec différents locuteurs (masculins et féminins). Nous avons calculé la différence moyenne absolue et l'écart type de suivi estimé par notre méthode en prenant les signaux étiquetés issus de la base élaborée comme référence tout en mettant l'utilité de notre base de données étiquetées manuellement. Nous avons tout d'abord fait des tests sur la voyelle /a/ précédée chaque fois d'une consonne. Ensuite, nous avons testé l'algorithme sur les différentes voyelles courtes et longues. Les résultats de suivi ont été ensuite comparés à ceux des méthodes crêtes de Fourier et d'analyse LPC mise en œuvre dans le logiciel (Praat).

D'après les résultats obtenus à la suite des tests sur la voyelle /a/ précédée par les différentes classes de consonnes, nous avons constaté que le suivi fait par la méthode ondelette avec cmor est globalement meilleur que celui des autres méthodes Praat et Fourier. Cette méthode donne donc un suivi de formants (F1, F2 et F3) pertinent et plus proche de suivi référence. Les résultats des méthodes Fourier et ondelette sont très proches dans certains cas puisque toutes les deux présentent moins d'erreurs que la méthode Praat.

D'après les résultats des valeurs de différence moyenne absolue et l'écart type, nous avons obtenu un suivi correct pratiquement par les trois méthodes et nous avons vérifié cela pour les cinq locuteurs masculins. Le suivi des trois formants est pertinent pour la plupart des cas. En revanche dans le cas où la voyelle est précédée par la consonne Tap /r/, on remarque que le suivi est bon pour F1 et F2 par les trois méthodes et pour les cinq locuteurs mais il y a des erreurs concentrées sur F3. De même dans les cas où la voyelle /a/ est précédée par les consonnes /m/ et /q/, le suivi est pertinent pour tous les locuteurs avec la méthode par ondelette mais entaché de quelques erreurs pour les deux autres méthodes, les erreurs apparaissent soit pour F1, F2 ou F3.

En faisant les mêmes tests avec les locutrices féminins, nous avons constaté que globalement le suivi de formants est correct pour les trois méthodes et pour les trois locutrices W3, W4 et W5. Le suivi des trois formants est pertinent pour la plupart des cas, sauf dans le cas où la voyelle est précédée par les consonnes /r/ et /d/, nous avons remarqué qu'on a toujours un bon suivi de F1 pour les trois méthodes même s'il y a parfois certaines erreurs locales surtout sur F3 et parfois sur F2. Cette observation vaut pour les trois locutrices.

Dans le cas des tests sur les voyelles longues et courtes avec les locuteurs masculins, nous avons noté d'après les résultats que les erreurs se manifestent le plus souvent sur F3 pour les voyelles /i/ et /I/ et spécialement pour les voyelles /u/ et /U/. Cette observation vaut pour les trois méthodes quelque soit le locuteur. Nous avons constaté aussi que les résultats des méthodes ondelette et Fourier ont été souvent proches dans certains cas.

Dans le cas des locutrices, les résultats ont montré que le suivi de F1, F2 et F3 est correct pour les voyelles /a/ et A/. Cela vaut pour les trois méthodes et les trois locutrices, à l'exception de la méthode LPC via Praat pour la locutrice W3 et les formants F2 et F3.

Dans le cas des voyelles /i/ et /u/, nous avons constaté que le suivi est correct pour F1 et F2 et qu'il y a quelques erreurs sur F3 pour les trois méthodes et avec les trois locutrices. En revanche, la méthode de Praat présente quant à elle des erreurs de suivi sur F2 et F3.

Dans le cas des voyelles longues /I/ et /U/, nous avons remarqué qu'on a des erreurs concernent les trois méthodes et pour les trois locutrices et essentiellement sur F2 et F3.

Finalement, nous avons testé les deux nouvelles approches de suivi, c'est-à-dire la première qui est basée sur la détection des crêtes d'ondelette utilisant le calcul de centre de gravité et la deuxième basée sur la détection des crêtes de Fourier en utilisant la programmation dynamique combinée avec le filtrage de Kalman, sur la parole continue en faisant des tests sur différentes phrases et avec différents locuteurs (masculins et féminins) et les résultats de suivi ont été comparés à ceux de la méthode par crêtes de Fourier en utilisant le calcul de centre de gravité, de l'analyse LPC combinée à des bancs de filtres de Mustafa Kamran (Mustafa.K 2003) (Mustafa.K 2006) et de l'analyse LPC dans le logiciel Praat.

D'après les résultats de ces différents tests, nous avons constaté que dans le cas des locuteurs masculins, les résultats des deux nouvelles approches sont notamment meilleurs que ceux de la méthode LPC de Mustafa Kamran et ceux de Praat même si elles présentent

souvent quelques erreurs sur F3. Elles sont aussi très proches de la méthode par détection de crêtes de Fourier utilisant le calcul de centre de gravité.

Les résultats obtenus dans le cas des locutrices féminins confirment la tendance observée sur les locuteurs. En effet les deux nouvelles approches de suivi donnent de meilleurs résultats que celle de Mustafa Kamran qui présente des erreurs très élevées qui dépasse les 200 Hz, même si globalement les résultats sont moins bons que pour les locuteurs masculins.

Parmi les erreurs que nous avons rencontrées fréquemment pendant ces différents tests sont celles qui sont concentrées surtout sur F3 pour les deux nouvelles approches. Ces erreurs sont généralement à cause soit de la faible énergie au niveau du spectrogramme ou du scalogramme, c'est-à-dire manque de fréquences candidates, soit à cause de l'effet de coarticulation dans la parole continue qui engendre la présence de faux pics et c'est ce qui mène à un faux suivi.

Nos futurs travaux consistent à améliorer ces deux approches par le fait d'utiliser les modèles HDM que Li Deng a suggéré dans sa méthode de suivi de formants pour compenser la probabilité d'avoir des fréquences candidates manquantes pendant le suivi et dans ce que nous serons plus dépendant uniquement de l'énergie du spectrogramme ou du scalogramme pour fournir l'ensemble des fréquences candidates pour chaque formant.

Pour notre nouvelle approche basée sur la programmation dynamique, nous proposons d'introduire dans un futur travail d'autres contraintes de continuité pour avoir un meilleur suivi surtout au niveau des transitions rapides entre les différents phonèmes.

Nous proposons aussi d'implémenter prochainement une nouvelle approche de suivi de formants basées sur les HMM en faisant l'apprentissage d'une petite base de test fondée sur un ensemble de signaux de notre base étiquetés à la main et introduire la dépendance du concept phonémique.

Bibliographie

Abdelatty Ali.A M, S. V. (2000). Auditory based speech processing based on the average localized synchrony detection. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), , Vol 3*, pp. pp.1623–1626.

Abdelatty Ali.A M, S. V. (2002). Robust auditory-based speech processing using the average localized synchrony detection. *Journal IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING , , Vol 10* (5), pp.279-292.

Acero.A. (1999). Formant analysis and synthesis using markov models. *Proceeding of ICASSP Microsoft Research.* Washington, USA.

Alessandro.C, D. (1992). Représentations temps-fréquence du signal de parole. *Traitement du signal , , Vol 9* (2), pp.153-173.

Ali.JA.M.A, S. M. (2002). Robust auditory based processing using the average localized synchrony detection. *IEEE Transactions on Speech and Audio processing* .

Álvarez.A, M. N. (1997). Continuous Formant-Tracking Applied to Visual Representations of the Speech and Speech Recognition. *Proceedings of EUROSPEECH'97.* Rhodes, Greece.

Anusuya.M.A, K. (2011). Front end analysis of speech recognition: a review. *Journal Int J Speech Technol , , Vol 14*, pp.99-145.

Arabic, W. (2011). *Apprendre L'arabe:les voyelles arabes.* Récupéré sur http://www.webarabic.com/portail/apprendre/index.php?rub=ecrire&p...

Bazzi.I, A. D. (2003). An Exception maximization approach for formant tracking using a parameter free non linear predictor. *Proceedings of ICASSP Microsoft Reasearch.*

Boudraa.M, B. G. (2000). Twenty lists of ten arabic sentences for assessment. *Journal ACUSTICA , , Vol 86*, pp.870-882.

Boukadida.F. (2006). *Etude de la prosodie pour un système de synthèse de la parole arabe standard à partir du texte.* Thèse, ENIT, Tunis.

Braham.A. (1997). *An acoustic study of temporal oraganization in arabic speific to tunisian speakers.* Thèse (writen in arabic), Université Manouba, Tunis.

Bruce. I C, K. N. (2002). Robust formant tracking in noise. *Journal Electrical and Electronic Engineering , , Vol 1*, pp.281-284.

Calliope. (1989). *La parole et son traitement automatique* (éd. CENT-ENST). Masson.

Chaari.S. (2005). *Suivi de formants par détection des crêtes d'ondelette.* Mastère, ENIT.

Châari.S, O. E. (2006). Wavelet ridge track Interpretation in terms of formants. *Proceedings of INTERSPEECH*, (pp. pp. 1017-1020). Pennsylvania, USA.

Chaplais.F. (2001). *A wavelet tour of signal processing by Stéphane Mallat*. Récupéré sur http://cas.ensmp.fr/~chaplais/Wavtour_présentation/

Charpentier, F. (1986). Pitch detection using the short-term phase spectrum . *Pceceedings of ICASSP*. Tokyo.

Chen.B, L. C. (2004). Formant frequency estimation in noise. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) , , Vol 5*.

cmorwav. Récupéré sur http://www.mathworks.com/help/toolbox/wavelet/ref/cmorwavf.html

Combescure.P. (1981). 20 listes de 10 phrases phonétiquement équilibrées. *Journal ACUSTICA* , pp.34-38.

Daoudi.K. (2004). *State of the art in speech and audio processing.* INRIA-Parole, France.

Darch.J, M. S. (2005). Predicting formant frequencies from MFCC Vectors. *Proceeding of ICASSP* , (pp. pp.941-944). Norwich.

Delsuc.M.A.*Traitement de signal.* Récupéré sur www.cbs.cnrs.fr/MAJ/FORMATIONS/COURS/RMN/cours/MAD/signal-R_eacute.html

Deng.Li. (2006). A database of vocal tract resonance trajectories for research of speech precessing. *Proceedings of ICASSP*, *,Vol 1*, pp. pp.369-372.

Deng.Li, B. A. (2006). Tracking Vocal Tract Resonances Using a Quantized Nonlinear Function Embedded in a Temporal Constraint. *Journal IEEE Trans. on Audio, Speech and Language Processing , , Vol 14* (2), pp. 425-434.

Deng.Li, L. A. (2004). A structured speech model with continous hidden dynamics and prediction residual training for tracking vocal tract resonances. *Proceeding of ICASSP*, (pp. pp.557-560).

Depalle.Ph, G. (1992). *Analyse des signaux sonores en termes de partiels et de bruit. Extraction automatique des trajets fréquentiels par des modèles de Markov Cachés.* Rapport de stage DEA Automatique et Traitement de Signal, IRCAM, Université Paris Sud.

Deppalle.Ph, G. R. (1993). Tracking of pratials for auditive sound synthesis using hidden Markov models. *Proceedings of ICASSP'93 IEEE Inetrnational Conference*, *,Vol 1*, pp. ,pp. 225-228.

Dugand.P. (1999). Phonétique et phonologie du Français. CEFISEM Nancy - Metz.

Dutoit.T. (2000). Introduction au traitement automatique de la parole. (Première Edition).

fbspwav. Récupéré sur http://www.mathworks.com/help/toolbox/wavelet/ref/fbspwavf.html

Gang.L, H. (2010). Developments of the research of the formant tracking algorithm. *Journal Computer and Information Science , ,Vol 3* (1), pp.68-71.

Gargouri.D. (2010). *Contribution à l'estimation et à la poursuite des trajectoires de formants de parole.* Thèse, Ecole Nationale d'Ingénieurs de Sfax, Sfax, Tunisie.

Ghazali.S. (1977). *Consonants and backing coarticulationin arabic.* Thèse, Université de Texas, Austin.

Gläser.C, H. J. (2010). Combining auditory preprocessing and bayesian estimation for robust formant tracking. *Journal IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING , ,Vol 18* (2), pp.224-236.

Haton.J.P, C. ,. (2006). *Reconnaissance automatique de la parole.* (Dunod, Éd.)

Introduction à la phonétique. Récupéré sur aix1.uottawa.ca/~hknoerr/MunotNeve115119.pdf

Jemâa.I. (2007). *Suivi de formants par détection des crêtes de Fourier.* Mastère, ENIT.

Knoerr.H. (2011). Récupéré sur aix1.uottawa.ca/~hknoerr/MunotNeve115119.pdf

Kopec.G. (1986). Formant Tracking Using Hidden Markov Models and Vector Quantization. *Journal IEEE Transactions on Acoustics, Speech, and Signal processing , , Vol 4* (34), pp. 709-729.

Kumaresan.R, A. C. (2005). Adaptive filter banks inspired by the auditory system for speech feature extraction. *Proceeding of ICASSP, , Vol 1*, pp. pp.905-908.

Laprie.Y. (2004). A concurrent curve strategy for formant tracking. *Proceedings of INTERSPEECH.*

Laprie.Y. (2009). Analyse spectrale de la parole.

Lee.M, S. M. (2005). Formant trcking using context dependent phonemic information. *IEEE Transactions On Audio, Speech and Language Processing , , Vol 13* (5), pp.741-750.

Léothaud.G. (2005). Théorie de la phonation.

Linguistique UNIL. (2011). *Caractéristiques acoustiques des différentes réalisations.* Récupéré sur http://www.unil.ch/ling/page13432.html

Mallat.S. (2000). *Une exploration des signaux en ondelettes* . Ecole Plytechnique.

Manisha.Miss, A. (2012). Theoretical Survey of the Formant Tracking Algorithm. *IJCA Proceedings on National Conference on Innovative Paradigms in Engineering & Technology (NCIPET , , Vol ncipet* (2), pp.14-17.

Manocha.S, E.-W. (2005). Knowledge-based formant tracking with confidence measure using dynamic programming. *Journal Acoustical Society of America , ,Vol 118* (3), pp. 1930-1930.

Marchal.A. (2007). *La production de la parole.* (Lavoisier, Éd.) Hermès.

Markel.J D, G. H. (1976). *Linear prediction of speech.* Springer-Verlag, New York, USA.

McCandless.S.S. (1974). An algorithm for automatic formant extraction using linear preditcion spectra. *Journal IEEE on Acoustics, Speech and Signal Processing , , Vol 22* (2), pp.135-141.

Metz.S W, H. A. (1991). Auditory modelling applied to formant tracking of noise corrupted speech. *Proceedings of the International Conference on Industrial Electronics, Control and Instrumentation*, *, Vol 3*, pp. pp.2120-2124.

Mustafa.K. (2003). *Robust formant tracking for continuous speech with speaker.* Thèse, Dept. Elect. and Comp. Eng. McMaster Univ, Hamilton, Canada.

Mustafa.K, B. C. (2006). Robust formant tracking for continuous speech with speaker variability. *Journal IEEE Transactions On Audio Speech And Language Processing , , Vol 14* (2), pp.435-444.

Mustafa.K, B. (2006). Robust formant tracking for continuous speech. *IEEE Transactions on Speech and Audio Processing* .

Newman.D. (s.d.). The phonetics of arabic. *Arabic Phonetics : Sound Descriptions* . Récupéré sur www.dur.ac.uk/daniel.newman/phon5.pdf

Newman.D. THE PHONETICS OF ARABIC. *Arabic Phonetics : Sound Descriptions* , (pp. pp.1-6).

Nguyen.N. (2001). Rôle de la coarticulation dans la reconnaissance des mots. *Année Psychologie*, *, Vol 101*, pp. pp.125-154.

O'Shaughnessy.D. (2008). Formant estimation and tracking. *Bouk Chapter in Springer Handbook of Speech Processing* , pp.213-228.

Ouni.K, L. E. (2001). Formant estimation using gammachirp filterbank. *Proceedings of INTERSPEECH.*

Özbek.I.Y, D. (2006). Tracking of visible vocal tract resonances (VVTR) based on kalman filtering. *Proceedings of INTERSPEECH.*

Özbek.Y.I, D. (2008). Vocal tract resonances tracking based on voiced and unvoiced speech classification using dynamic programming and fixed interval kalman smoother. *Journal IEEE Transactions on ICASSP* , pp.4217-4220.

Papadakis.N. (2007). *Assimilation de données images : application au suivi de courbes et de champs de vecteurs.* thèse, IFSIC, Université de Rennes 1.

Pinquier.J. (2004). *Indexation Sonore : Recherche de Composantes Primaires pour une Structuration Audio-Visuelle.* Thèse.

Plante.F, A. (1994). Formant tracking: A comparison of several parametric methods. *Proceedings of the Institute of Acoustic*, , *Vol 16*, pp. pp.293-300.

Praat. Récupéré sur http://www.praat.org/ .

Rachedi.J. (2005). *Reconnaissance et classification de phonèmes.* Mastère, Laboratoire IRCAM, Paris.

Rekhis.O. (2009). *Elaboration et étiquetage phonétique et formantique d'un corpus en langue arabe. Application au suivi de formants.* Mastère, Ecole Nationale d'Ingénieurs de Tunis.

Richards.H.B, B. (1999). The HDM: A segmental hidden dynamic model of coarticulation. *Proceeding of ICASSP*, (pp. pp.357-360).

Rudoy.D, S. W. (2007). Conditionally linear Gaussian models for estimating vocal tract resonances. *Proceedings of INTERSPEECH (ISCA)*, (pp. pp.526-529).

Saidane.T, Z. B. (2004). La transcription orthographique phonétique de la langue arabe. *Proceedings of RECITAL.* Fès.

SAMPA. (s.d.). Récupéré sur http://www.phon.ucl.ac.uk/sampa/home.htm/.

Sandeep.M, E.-W. Y. (2005). Knowledge-based formant tracking with confidence measure using dynamic programming. *Journal Acoustical Society of America Journal , , Vol 118* (3), pp.1930-1930.

Satori.H, H. C. (2007). Système de Reconnaissance Automatique de l'arabe basé sur CMUSphinx .

Schafer.R W, R. R. (1970). System for automatic formant analysis of voiced speech. *Journal of the Acoustical Society of America , , Vol 47* (2), pp.634-648.

Secrest.B, D. (1983). An integrated pitch tracking algorithm for speech systems. *Proceedings of ICASSP*, (pp. pp. 1352-1355).

Seneff.S. (1988). A joint Synchrony/Mean rate model of auditory speech processing. *Journal Pkonrtics , , Vol 16*, pp.55-76.

Shahidur Rahman.M, S. (2005). Formant frequency estimation of high-pitched speech by homomorphic prediction. *Journal Acoustical Science and Technology , , Vol 26* (6), pp.502-510.

Shamma.S. (1988). The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives. *Journal Phonetics , , Vol 16*, pp.77-91.

Shanwav. Récupéré sur http://www.mathworks.com/help/toolbox/wavelet/ref/shanwavf.html

Sharma.M, M. (1996). Blind speech segmentation: Automatic segmentation os speech without linguistic knowledge. *Proceedings of ICSLP*, (pp. pp.1237-1240).

Sjölander.K. (2003). An HMM-based system for automatic segmentation and alignment of speech. *In Proceedings of Fonetik*, (pp. pp. 93-96).

Snell.Roy C, M. (1993). Formant location frop LPC analysis data. *Journal IEEE Transactions on Speech and Audio Processing , , Vol 1* (2), pp.129-134.

Sun.X, D. (1995). Robust estimation of spectral center-of-gravity trajectories using mixture spline models. *Proceedings of EUROSPEECH.*

Sundaram.N, Y. S. (2003). Usable speech detection using linear predictive analysis model based approach . *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)* , (pp. pp.231-235). Awaji Island, Japan.

Talkin.D. (1987). Speech formant trajectory estimation using dynamic programming with modulated transition costs. *Journal of the Acoustical Society of America , , Vol 82*.

Toledano.T.D, G. G. (2003). Automatic phonetic segmentation. *IEEE Transactions On Audio, Speech and Language Processing , , Vol 11* (6), pp.617-625.

Toledano.T.D, V. G. (2006). Initialization, training and context dependency in HMM based tracking. *IEEE Transactions On Audion Speech and Language Processing , , Vol 14* (2), pp.511-523.

Vargas.J, M. (2008). Cascade prediction filters with adaptive zeros to track the time-varying resonances of the vocal tract. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING , , Vol 16* (1).

Wavesurfer. Récupéré sur http://sourceforge.net/projects/wavesurfer/

Winsnoori. Récupéré sur http://www.loria.fr/~laprie/WinSnoori/.

Xia.K, E.-W. (2000). A new strategy of formant tracking using dynamic programming. *Proceeding of Int.Conf.Spoken Lang.Proce.*

Xwaves 5.3.1. Récupéré sur Entropic Research Laboratory Inc.

Youssef.A, E. (2004). An Arabic TTS System Based on the IBM Trainable Speech Synthesizer. *JEP–TALN* . Fès.

Yves.L. (2004). A concurrent curve strategy for formant tracking. *INTERSPEECH.*

Zaki.A. (2004). *Modélisation de la prosodie pour la synthèse de la parole arabe standard à partir du texte.* Thèse, ESPI, Université Bordeaux.

Zheng.Y, J. (2004). Formant tracking by mixture state particle filter. *Proceeding of ICASSP, , Vol 1*, pp. pp.981-984.

Annexe A: API

Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.
/f/	/ف/	/ ʔ /	/ء/	/l/	/ل/	/ ʃ /	/ش/	/ tˤ/	/ط/
/s/	/س/	/q/	/ق/	/r/	/ر/	/ θ /	/ث/	/t/	/ت/
/ sˤ/	/ص/	/ dˤ/	/ض/	/m/	/م/	/x/	/خ/	/ ʕ /	/ع/
/z/	/ ز/	/d/	/د/	/n/	/ن/	/ ð/	/ذ/	/ dʒ /	/ج/
/h/	/ه/	/b/	/ب/	/w/	/و/	/ ðˤ/	/ظ/		
/ ħ /	/ح/	/k/	/ك/	/j/	/ي/	/ ɣ /	/غ/		

Correspondance graphème phonème de la langue arabe standard selon l'alphabet phonétique internationale (API)

Annexe B: Constitution du Corpus

Liste 1

• ʔaħfadˁu mina lʔardˁi • Plus conservateur que la terre	أَحْفَظُ مِنَ الأَرضِ
• ʔajna lmusa:firu:na ? • Où sont les voyageurs ?	أَيْنَ المُسَافِرُونَ ؟
• la: lam jastamtiʕ biθamariha: • Non! Il n'a pas joui de ses fruits!	لاَ لَمْ يَسْتَمْتِعْ بِثَمْرِهَا.
• sajuʔði:him zama:nuna: • Notre temps les blessera.	سَيُؤْذِيهُمْ زَمَانُنَا.
• Kuntu qudwatan lahum • J'ai été un exemple pour eux.	كُنْتَ قُدْوَةً لَهُمْ.
• ʔa:zara sˁa:ʔiman • Il récompense un homme à jeun	آزَرَ صَائِمًا.
• Ka:la wa ɣabatˁa lkabʃa • Il a mesuré et renforcé les moutons.	كَالَ وَغَبَطَ الكَبْشَ.
• Hal laðaʕathu bi qawlin ? • L'as-t-elle blessé avec les mots?	هَلْ لَذَعَتْهُ بِقَوْلٍ؟
• ʕarafa wa:lijan wa qa:ʔidan • Il a connu un gouverneur et un commandant.	عَرَفَ وَالِيًا وَقَائِدًا.
• xala: ba:luna: minkuma: • On n'a pas fait attention à vous.	خَلاَ بَالُنَا مِنْكُمَا

Liste 2

• la: lan juði:ʕa lxabara • Non! Il ne répandra pas la nouvelle.	لاَ لَنْ يُذِيعَ الخَبَرَ.
• ʔakmil bilʔisl:mi ris:lataka • Finis ta lettre avec l'Islam.	أَكْمِلْ بِالإِسْلاَمِ رِسَالَتِكَ.
• saqatˁat ʔibratun • Une aiguille est tombée.	سَقَطَتْ إِبْرَةٌ.
• Man lan jantafiʕ ?	مَنْ لَنْ يَنْتَفِعْ؟

• Qui ne profitera pas ?	
• ɣafala ʕan dˤaħaka:tiha: • Il n'a pas pris ses rires en considération.	غَفَلَ عَنْ ضَحَكَاتِه
• wa lima:ða naʃafa ma:luhum ? • Pourquoi ont-ils perdu leur fortune?	وَلِمَاذَا نَشَفَ مَالُهُمْ؟
• ʔayna zawa:ya:na: wa qa:nu:nuna:? • Où sont nos coins et nos harps?	أين زَوَايَانَا و قَانُونُنَا؟
• sˤa:da lmawru:θu modlijan • L'héritier a chassé un hérisson	صَادَ المَوْرُوثُ مُدْلِجًا.
• nabiha ʔabaʔukum • Tes ancêtres ont eu une bonne réputation.	نَبِهَ آبَاؤُكُمْ.
• qum wa ʔaðˤhirhu • Lève-toi et montre-le	قُمْ وَأَظْهِرْهُ

Liste 3

• bilwa:lidajni iħsa:nan • Il faut prendre soin de ses parents	بِالوَالِدَيْنِ إحْسَانًا.
• ʔistaqim kama: ʔumirta • Sois droit comme on te l'a demandé	إسْتَقِمْ كَمَا أُمِرْتَ.
• ka:na lʔaklu laði:ðan • Le repas a été délicieux	كَانَ الأَكْلُ لَذِيذًا.
• hal ha:ra ? • A-t-il tombé à la renverse ?	هَلْ هَارَ؟
• ʔasaru:na: bimunʕatafin • Ils nous ont capturés dans un virage.	أسَرُونَا بِمُنْعَطَفٍ.
• jamaʕa lmawza wa xala: • Il a récolté les bananes et il est parti	جَمَعَ المَوْزَ وَخَلَى.
• dˤamintu ʃaɣafakum • J'ai assuré votre passion	ضَمِنْتُ شَغَفَكُمْ.
• warima: falan juqa:tila: • Ils se sont blessés, alors ils ne se battront pas	وَرِمَا فَلَنْ يُقَاتِلَا.
• hija huna: laqad ʔa:bat • Elle est ici et elle était pieuse	هِيَ هُنَا لَقَدْ آبَتْ.
• ʔabsara θuʔba:nan wa lam jaðˤlimhu • Il a vu un serpent et n' a pas l'opprimer.	أبْصَرَ ثُعْبَانًا وَلَمْ يَظْلِمْهُ

Liste 4

• tˤafaħa lkajlu • Il y en a marre	طَفَحَ الكَيْلُ.
• laqad ka:na musa:liman wa qutila • Il était pacifique et il a été tué	لَقَدْ كَانَ مُسَالِمًا وَقُتِلَ.
• La:na wa lam jakun ʃarisan • Il était docile et n'était pas féroce	لاَنَ وَلَمْ يَكُنْ شَرِسًا.
• dˤamina θawratahum • Il a assuré leur révolution	ضَمِنَ ثَوْرَتَهُمْ.
• lan juna:siru:hum ɣadan • Ils ne vont pas les soutenir demain	لَنْ يُنَاصِرُوهُمْ غَدًا.
• ʔaðˤfir bima: ʔaxaðat • Il a triomphé avec ce qu'elle a eu	أَظْفِرَ بِمَا أَخَذَتْ.
• ʔatuʔði:ha: biʔa:la:mihim ? • Tu la blesses avec leurs peines?	أَتُؤْذِيهَا بِآلامِهِمْ؟
• ʔistalzama lʕafanu waʕjakuma: • Le putride domine votre conscience	إِسْتَلْزَمَ العَفَنُ وَعْيَكُمَا.
• qabalna: wabran • Nous avons fait face à un fennec	قَابَلْنَا وَبْرًا.
• hal dʒaːʕa abun ? • Le père a-t-il faim ?	هَلْ جَاعَ أَبٌ؟

Liste 5

• ʔaxfaqa mutaʔa:miru:na • Les conspirateurs n'ont pas réussi	أَخْفَقَ مُتَآمِرُونَ.
• ʔistaɣfir liðanbika • Demande le pardon pour ton péché.	إِسْتَغْفِرْ لِذَنْبِكَ.
• ha ana: ði: ʔunsitu li jaqul ! • Je suis en train d'écouter, qu'il parle !	هَا أَنَا ذِي أَنْصِتُ لِيَقُلْ
• qadana: wa lam jaðˤtˤahidkum • Ils nous ont commandés et ils ne vous ont pas méprisé.	قَادَنَا وَلَمْ يَضْطَهِدْكُمْ.
• ma labisa θawban • Il n'a jamais mis de vêtements.	مَا لَبِسَ ثَوْبًا.
• ka:na minhum fi: ðˤuluma:tin • Il était dans l'obscurité en ce qui les concerne.	كَانَ مِنْهُمْ فِي ظُلُمَاتٍ.
• ʔakrimhu wa ʕmal • Sois généreus avec lui et travaille	أَكْرِمْهُ وَأَعْمَلْ
• la: ja zeinabu ʃarama ħablan • No Zeineb ! il déroule une corde	لا يَا زَيْنَبُ شَرَمَ حَبْلاً

• ɣala: wa dʒaza: lʔinsa:nu • l'homme est de la supériorité et de l'oppression.	غَلاَ وَجَازَ الإِنْسَانُ.
• wa hal ʕalat ? • S'est-t-elle lever ?	وَهَلْ عَلَتْ؟

Liste 6

• xalaqa lʔinsa:na min nutˁfatin • Dieu a crée l'humain d'une cellule.	خَلَقَ الإِنْسَان مِنْ نُطْفَةٍ.
• yuqa:miru:na bilma:li •	يُقَامِرُونَ بِالمَالِ.
• la: lam jastabiħ ɣadrahum	لاَ لَمْ يَسْتَبِحْ غَدْرَهُمْ.
• na:ðˁara lmuði:ʕa	نَاظَرَ المُذِيعَ.
• la: takun ʃarisan • Ne sois pas féroce	لا تَكُنْ شَرِسًا.
• ʔa:θama lʔabna:ʔa wa zawdʒaha:	آثَمَ الأَبْنَاءَ وَزَوْجَهَا.
• dˁaʔula wa lam jarkaʕ lil wa:qifi	ضَوُلَ وَلَمْ يَرْكَعْ لِلْوَاقِفِ.
• fasadat ða:tu bajnihim	فَسَدَتْ ذَاتُ بَيْنِهِمْ.
• ka:na sˁa:ʔiman • Il a jeuné	كَانَ صَائِمًا.
• wakuʕa ʔabu:huma: •	وَكُعَ أَبُوهُمَا.

Liste 7

• ʔasˁa:bakum biqaði:fatin • Il vous a touchés avec une bombe.	أَصَابَكُمْ بِقَذِيفَةٍ.
• la:θa mudminun •	لاثَ مُدْمِنٌ.
• ʔaxaða ʔidʒa:zatan • Il a pris un congé	أَخَذَ إِجَازَةً.
• lan wa lan jana:laha: • Il ne l' aura jamais.	لَنْ وَلَنْ يَنَالَهَا.
• ʔaryamatka dˁaru:ratun • Il a été obligé.	أَرْغَمَتْكَ ضَرُورَةٌ.
• sa:ħa lma:ʔu liðˁamʔa:na	سَاحَ المَاءُ لِظَمْآنَ.

• wa stabʃaʕa sawtahum •	وَاسْتَبْشَعَ سَوْطَهُمْ.
• rafaʕa jadyhi lilba:riʔi fi: mana:min • Il a prié Dieu dans le rêve.	رَفَعَ يَدَيْهِ لِلْبَارِئِ فِي مَنَامٍ.
• hal ka:na juqa:bilukuma: ? • Il vous a rencontrés?	هَلْ كَانَ يُقَابِلُكُمَا؟
• wala: wa lam naʕqilhu • Il est parti et on ne l'a pas reconnu	وَلَى وَلَمْ نَعْقِلْهُ.

Liste 8

• qa:da ldʒajʃa • Il a commandé l'armé	قَادَ الجَيْشَ.
• sajusqitˁu muʔa:marataka •	سَيُسْقِطُ مُؤَامَرَتَكَ.
• baʕaθat naði:ran •	بَعَثَتْ نَذِيرًا.
• ʔiðhab bi ʔama:nin • Bonne route	إِذْهَبْ بِأَمَانٍ
• Kana: fi: ðˁuluma:tin wa lam jarħal • Il était dans l'obscurité et il n'est pas parti.	كَانَ فِي ظُلُمَاتٍ وَلَمْ يَرْحَلْ.
• sˁanaʕa midfaʔatan wa miʕzafan • Il a fabriqué un chauffage et un instrument de musique.	صَنَعَ مِدْفَأَةً وَمِعْزَفًا.
• law la: ʔan maridˁna: laxasiru: • Si on n'était pas tombé malade, ils auraient perdu.	لَوْلا أَنْ مَرِضْنَا لَخَسِرُوا.
• lam jaktumhu • Il ne l'a pas caché.	لَمْ يَكْتُمْهُ.
• wala:ʔuha: lilqalbi wa liɣaba:ʔihim • Sa fidélité à son cœur et à leur stupidité.	وَلاؤُهَا لِلْقَلْبِ وَلِغَبَائِهِمْ.
• kun huna: • Sois là.	كُنْ هُنَا

Liste 9

• ma:ða: juði:bu ? • Qu'est ce qu'il fait fondre.	مَاذَا يُذِيبُ؟
• zamina wa lam jantafiʕ bibalsamihim	زَمِنَ وَلَمْ يَنْتَفِعْ بِبَلْسَمِهِمْ.
• ʔabrim lana: ʔumu:rahum	أَبْرِمْ لَنَا أُمُورَهُمْ.
• wahat faʔa:wa:ha: • Elle est devenue faible, alors il a logée.	وَهَتْ فَآوَاهَا.
• la: tuka:bid lan jaðˤlima sa:ʔiqan •	لا تُكَابِدْ لَنْ يَظْلِمْ سَائِقًا.
• daxalu: lmuru:dʒa	دَخَلُوا المُرُوجَ.
• kul ʕajʃaka wa staħdˤir lahuma: qa:nu:nan	كُلْ عَيْشَكَ وَاسْتَحْضِرْ لَهُمَا قَانُونًا.
• kun sˤa:ʔiɣan • Ouvres bien tes oreilles.	كُنْ صَائِغًا.
• ʔawraθa la:qitˤan •	أَوْرَثَ لَا قِطًا.
• ʔalifati lʕana:na •	أَلِفَتِ العَنَانَ.

Liste 10

• wajlun lildʒa:ʔiri ! •	وَيْلٌ لِلْجَائِر
• ʔalam jastansˁir bilqa:tili ? • A-t-il donné raison au meurtrier ?	أَلَمْ يَسْتَنْصِرْ بِالقَاتِلِ؟
• ʔalam jakun maʕru:fan wa ʃaɣala muði:ʕan ? • N'a-t-il pas été connu et a travaillé dans une station radio ?	أَلَمْ يَكُنْ مَعْرُوفًا وَشَغَلَ مُذِيعًا؟
• haða: waraħala	هَذَى وَرَحَلَ.
• wantafaʔa dˁawʔuhum • Leur lumière s'est éteinte.	وَانْطَفَأَ ضَوْؤُهُمْ.
• ka:nat θaknatuhum fi: sala:min	كَانَتْ ثُكْنَتَهُمْ فِي سَلامٍ.
• qad ʔa:ba lkalbu	قَدْ آبَ الكَلْبُ.
• la:zama: ðˁabja:na:	لازَمَا ظَبْيَانَا.
• ma: xadaʕa namiraha:	مَا خَدَعَ نَمِرَهَا.
• nuhima bisa:ʔiqana:	نُهِمَ بِسَائِقِنَا.

Liste 11

• ma: qawluka fi: ðˁulmihim ?	مَا قَوْلُكَ فِي ظُلْمِهِمْ؟
• ʔistamiʕ lil ʔaða:ni	اسْتَمِعْ لِلأَذَانِ.
• tudrikuhum ɣadan bijama:matin	تُدْرِكهُمْ غَدًا بِيَمَامَةٍ.
• kun raði:lan	كُنْ رَذِيلاً.
• wadˁaʕa musˁtˁalaħan	وَضَعَ مُصْطَلَحًا.
• sajabʕaθu lʔabu ʔa:lahum	سَيَبْعَثُ الأَبُ آلَهُمْ.
• dʒala: wa halaka: bilbaʔsa:ʔi	جَالا وَهَلَكَا بِالبَأْسَاءِ.
• la:nat ʕuna:ha:	لَانَتْ عُنَاهَا.
• ʔajna na:ma wa:qifan ?	أَيْنَ نَامَ وَاقِفًا؟

• xazana lqa:ru:ta wa farʃan	خَزَنَ الْقَارُورَةَ وَفَرْشًا.

Liste 12

• ʔamtiʕna: binaɣamin	أَمْتِعْنَا بِنَعَمٍ.
• sajuʔði:hima: ʔiða: stabkajtukum	سَيُؤْذِيهِمَا إِذَا اسْتَبْقَيْتُكُمْ.
• Ka:na ljamanu fi: ʃaqa:ʔin	كَانَ اليَمَنُ فِي شَقَاءٍ.
• tˤalaba raqsˤatan mina lʕaru:si	طَلَبَ رَقْصَةً مِنَ العَرُوسِ.
• faʔa:θarathu zawdʒan	فَآثَرَتْهُ زَوْجًا.
• lam jakun mudrikan	لَمْ يَكُنْ مُدْرِكًا.
• ʕala: bilma:li wa ðˤalamahum	عَلَى بِالمَالِ وَظَلَمَهُمْ.
• na:walaha: lʔardˤi	نَاوَلَهَا الأَرْضِ.
• la:ħa wabi:lun	لاحَ وَبِيلٌ.
• ʔaxfi wa:lidaha:	أَخْفِ وَالِدَهَا.

Liste 13

• law lam jaʃ tˤubhum laquhirna:	لَوْ لَمْ يَشْطُبْهُمْ لَقَهِرْنَا.
• ɣafat fa ma:ta muxtaniqan	غَفَتْ فَمَاتَ مُخْتَنِقًا
• dʒa:ʔa bi dˤma:na:tin lil hurubi minka	جَاءَ بِضَمَانَاتٍ لِلْهُرُوبِ مِنْكَ.
• ʔajna naði:ratukum ?	أَيْنَ نَذِيرَتُكُمْ؟
• ʔaqama lʕadla wa a:zara ʔisman	أَقَامَ العَدْلَ وَآزَرَ اسْمًا.
• Jariθukum ba:lun	يَرِثْكُم بَالٌ.
• la: nansa lwaʕaðˤa wal wasa:ya	لا نَنْسَ الوَعَظَ وَالوَصَايَا.
• kafiʔ ħaða:mi	كَافِئِ حَذَامِ.
• ʕabada lʔila:ha	عَبَدَ الإِلَهَ.

• ʔin ha:sa:	إِنْ هَاسَا

Liste 14

• ʃaribta min ɣaru:fin	شَرِبْتُ مِنْ غَرُوفٍ.
• ðahaba ila: tilimsa:n	ذَهَبَ إِلَى تِلِمْسَانَ.
• wadˤaʕat ħamlaha:	وَضَعَتْ حَمْلَهَا.
• la lan juði:qakum min ʔaklina:	لا لَنْ يُذِيقَكُمْ مِنْ أَكْلِنَا.
• wa qubila ðˤulmuhum	وَقُبِلَ ظُلْمُهُمْ.
• qa:bilhu jawman ma: fi maka:nihima:	قَابِلْهُ يَوْمًا مَا فِي مَكَانِهِمَا.
• ʔaxtˤaʔta faʔa:θara sˤajdana:	أَخْطَأْتَ فَأَثَرَ صَيْدَنَا.
• ʔistadrakna: lħaba:nu	إِسْتَدْرَكْنَا الحَبَانُ.
• nawa: lʔaʕzalu	نَوَى الأَعْزَلُ.
• lasaʕa waryan	لَسَعَ وَرْيًا.

Liste 15

• ʔibnuka ʃaru:bun	إِبْنُكَ شَرُوبٌ.
• naxa wa lam jastaħsina ðˤulmukum	نَحَى وَلَمْ يستحسن ظُلْمَكُمْ.
• ha:wadatna: wa ka:faʔati lʕaði:ra	هَاوَدَتْنَا وَكَافَأَتِ العَذِيرَ.
• lan jula:misa waʕlan	لَنْ يُلَامِسَ وَعْلاً.
• naθura ʕalajhim jahi:mu:	نَثَرَ عَلَيْهِمْ يَهِيمُوا.
• dˤaminta fawzahom	ضَمِنْتُ فَوْزَهُمَا.
• ʔadbara lʔa:biqu	أَدْبَرَ الآبِقُ.
• ʔaxaðtana fi qaribin lil ma:li	أَخَذْتَنَا فِي قَارِبٍ لِلْمَالِ.
• natˤaqa sˤaʔimun	نَطَقَ صَائِمٌ.

ɣala: ʔakluhum	غَلا أَكْلُهُمْ.

Liste 16

• xasirat wala:ʔiman	خَسِرَتْ وَلائِمًا.
• ka:na mawtuha: sˁaʕban bisabi:likum	كَانَ مَوْتُهَا صَعْبًا بِسَبِيلِكُمْ.
• ħadˁara lwa ɣlu	حَضَرَ الوَغْلُ.
• lam judʒa:hid	لَمْ يُجَاهِدْ.
• ʔibnukum fa:ruqun wa jaðˁlimu	إِبْنُكُمْ فَارُوقٌ وَيَظْلِمُ.
• θaʔarta minhuma:	ثَأَرْتَ مِنْهُمَا.
• ʔasʕada lʔa:la bil qaði:fi	أَسْعَدَ الآلَ بِالقَذِيفِ.
• zaʕamna: an lan jaʃtarika	زَعَمْنَا أَنْ لَنْ يَشْتَرِكَ.
• huna la:tˁa lhajma:na	هُنَا لاطَ الهَيْمَانَ.
• qa:wala fi: miʔðanatin	قَاوَلَ فِي مِئْذَنَةٍ.

Liste 17

• Fahal ka:nat ħalabu ima:ratun	فَهَلْ كَانَتْ حَلَبُ إِمَارَةً .
• raʔajta qaði:fan	رَأَيْتُ قَذِيفًا.
• ʕala: wa ʔistaʕðˁama lʔinsa:nu	عَلا وَاسْتَعْظَمَ الإِنْسَانُ.
• lima:ða ʔaxtaʔa ʔa:damu ?	لِمَاذَا أَخْطَأَ آدَمُ؟
• saja ʃkurunana: fi: nadwatin	سَيَشْكُرُونَنَا فِي نَدْوَةٍ.
• man ɣadˁiba liʕilmihum ?	مَنْ غَضِبَ لِعَلَمِهِمْ؟
• qa:bilhum wa ʔi sˁbir	قَابِلْهُمْ وَاصْبِرْ.
• la: lam jobqi zawdʒaha:	لَا لَمْ يُبْقِ زَوْجَهَا.
• Wa ʔa:lun jariθu milkana:	وَأَلٌ يَرِثُ مَلِكُنَا.
• kon houna:	كُنْ هُنَا

Liste 18

• samiʕna: qaði:ʕatan	سَمِعْنَا قَذِيعَةً.
• ʔarsilhum liʔalija:ʔihim	أَرْسِلْهُمْ لأَوْلِيَائِهِمْ.
• ba:ʕa tˁablan faxaradʒa masruran	بَاعَ طَبْلًا فَخَرَجَ مَسْرُورًا.
• na:hadˁa lmodminu waħʃan	نَاهَضَ المُدْمِنُ وَحْشًا.
• la:janta man wafada ʔilajkum	لايَنْتَ مَنْ وَفَدَ إِلَيْكُمْ.
• qutila wariθun	قُتِلَ وَارِثٌ .
• Wa bima:ða la:zakuma ɣ:	وَبِمَاذَا لَا غَزَكُمَا؟
• ʔa sˁ/a:batha: bilkanafi	أَصَابَتْهَا بِالكَنَفِ.
• qa:ja ðˁa ʔa:lahum	قَايَظَ آلَهُمْ.
• natuna lʔaklu	نَتُنَ الأَكْلُ

Liste 19

• rafadˁa lfidjata fama: laha: min ʕawdatin	رَفَضَ الفِدْيَةَ فَمَا لَهَا مِنْ عَوْدَةٍ.
• ðahaba sˁiba:na	ذَهَبَ صِبَانَا.
• saʕjukum maqrunun bil ʕanaʔi	سَعْيُكُمْ قَرُونٌ بِالعَنَاءِ.
• ʔasqatˁat ʔindʒa:zakum	أَسْقَطَتْ إِنْجَازَكُمْ.
• ʔabliɣhu bil ʔamri	أَبْلِغْهُ بِالأَمِرِ.
• la:jana lajθan	لايَنَ لَيْثًا.
• xara ʃ tu lmirʔa:ta	خَرَشْتُ المِرْآةَ.
• waqa:ka lmaði:mu	وَقَاكَ المَذِيمُ
• ka:na wa:hinan wa la:ħ ðˁa waʔlan	كَانَ وَاهِنًا وَلاحَظَ وَألا.

• sa:la ma:luhum	سَالَ مَالُهُمْ.

Liste 20

• ħafitˁtu lqurʔa:na	حَفِظْتُ القُرْآنَ.
• qad nadˁbituhum	قَدْ نَضْبِطُهُمْ.
• saʕjuhum liʔabna:ʔihim	سَيُهُمْ لأَبْنَائِهِمْ.
• lan jastaʔminaha: ʔabadan	لَنْ يَسْتَأمِنَهَا أَبَدًا.
• lam jakuni lʔaklu laði:ðan	لَمْ يَكُنِ الأَكْلُ لَذِيذًا.
• ma: la:zama maɣru:ran wa ma: waraθa	مَا لازَمَ مَغْرُورًا وَمَا وَرثَ.
• ʕala: sˁaxabun bil dʒa:miʕati	عَلَا صَخَبٌ بِالجَامِعَةِ.
• halaka ʃawkun	هَلَكَ شَوْكٌ.
• sa:ra lqa:ʔimu	سَارَ القَائِمُ.
• waj linta fawa:fatna:!	وَيْ لِنْتَ فَوَافَتْنَا