Etude Comparative des Performances de Plusieurs Techniques de Détection de la Fréquence Fondamentale des Signaux Vocaux

F. Ykhlef [1], R. Amiar[1], S. Hecini[1], W. Benzaba[1], L. Bendaouia[1]

[1] Architectures des Systèmes et Multimédia
CDTA, Algérie
{ F. Ykhlef , R. Amiar, S. Hecini, W. Benzaba, L. Bendaouia}
ykhlef_faycal@yahoo.fr

Résumé. L'objectif de cet article est faire une étude comparative des performances de plusieurs méthodes de base d'extraction de la Fréquence Fondamentale des signaux vocaux prononcés par des locuteurs de différents sexes et âges (Femme, Homme et Enfant) en Arabe Standard. Quatre techniques particulières sont choisies, deux techniques temporelles, la MACC (Modified Autocorrelation with Center clipping), la C-AMDF (Clipping Average Magnitude Difference Function), et deux fréquentielles, la CEP (Cepstral Technic), et l'HPS (Harmonic Product Spectrum). Une représentation détaillée de ces techniques d'estimations et des méthodes de classifications employées est donnée dans cet article. L'évaluation des techniques est basée sur le calcul d'erreurs d'estimations.

Keywords: Fréquence Fondamentale, erreurs d'estimations, Arabe Standard.

1 Introduction

La fréquence la plus basse dans le signal de parole est la fréquence Fondamentale (F_0) dénommé « pitch ». Elle représente la fréquence de vibration des cordes vocales et caractérise les segments Voisés de la parole à l'intérieur desquels elle évolue lentement dans le temps. La plage de variation moyenne de cette fréquence varie d'un locuteur à l'autre en fonction de son âge et de son sexe. Elle s'étend approximativement de 80 à 200 Hz chez les hommes, de 150 à 450 Hz chez les femmes, et de 200 à 600 Hz chez les enfants [1].

Au cours des trente dernières années, un certain nombre d'algorithmes d'estimation de la F_0 ont été développés et rapportés. Ceci soulève la question évidente, pourquoi de nouveaux travaux sont toujours entrains d'être menés dans ce domaine?. Ainsi, aucun des nombreux algorithmes rapportés ne s'est avéré entièrement satisfaisant. Par conséquent, les chercheurs continuent à essayer d'obtenir des techniques améliorés pour l'estimation de la F_0. L'évaluation des algorithmes de détection de la F_0 est une opération importante pour les applications relatives au traitement de la parole [2].

On peut citer, l'analyse, la synthèse, le codage, la reconnaissance, la réverbération et les applications relatives à l'amélioration du confort d'écoute.

De ce fait, l'objective de ce travail est de procéder par une évaluation qualitative des algorithmes de base d'extraction de la F_0. Le choix est porté sur quatre techniques. Deux temporelles, à savoir, la MACC (Modified Autocorrelation with Center clipping), la C-AMDF (Clipping Average Magnitude Difference Function) et deux fréquentielles, la CEP (Cepstral Technic) et l'HPS (Harmonic Product Spectrum). Les critères d'évaluations sont basés sur le calcul d'erreurs d'estimations de la F_0 grossières et fines [3]. Un paramètre supplémentaire est ajouté au niveau de cette étude pour déterminer l'exactitude de ces dernières, il est nommé Paramètre d'Extraction Sans Erreur Commise (PESEC). Il caractérise les capacités théoriques de ces algorithmes à extraire le Fondamental à une erreur d'estimation exactement nulle.

L'évaluation des techniques de détection nécessite une base de données des sons spécifique à la langue traitée. Elle doit nécessairement contenir tout les classes des sons du langage ainsi que les dialectes utilisés [4]. Dans notre cas, on a préféré de se consacrer aux sons prononcés en Arabe Standard (AS). Du fait qu'il n'existe pas une base de données fiable pour cette langue, on a constitué un corpus modeste qui englobe tout les classes sonores de l'AS prononcés par des locuteurs de différents âges et sexes.

L'article est structuré en plusieurs parties. La première est réservée à la description des complexités d'extraction de la F_0 d'un signal vocal. La deuxième partie est consacrée à la présentation des méthodes d'extraction suivie en troisième partie par une description algorithmique des techniques implémentées. En quatrième partie, on présente les signaux de tests utilisés ainsi que la F_0 de référence. La cinquième partie traite les paramètres d'erreurs utilisées pour l'évaluation des techniques d'estimation.

La sixième partie est réservée à l'évaluation des performances des techniques par le calcul d'erreurs d'estimations suivi par une conclusion et des perspectives futures.

2 Complexités de détection de la F_0

La complexité d'évaluation du Fondamental est une tâche difficile pour de nombreuses raisons telles que la non-stationnarité du signal vocal, une certaines irrégularités dans l'excitation glottique ou encore une interaction avec le premier formant, la décision du voisement, la distinction entre les segments non voisée et les segments voisée à énergie réduite, la difficulté inhérente en définissant le début et la fin exacts de chaque période de F_0 durant les segments de la parole Voisée et dernièrement le doublement de période local [3]. C'est un type d'erreur qui affecte pratiquement toutes les méthodes d'estimation de la F_0.

3 Méthodes de détection de la F_0

D'après les travaux de Hess [5], les algorithmes de détection de pitch sont classées en trois groupes principaux : temporelles, spectrales et Hybrides.

Les méthodes temporelles permettent l'estimation de la F_0 avec des calcules très simples. Elles sont relativement peu couteuses en temps de calcul car elles nécessitent peu d'opérations arithmétiques de multiplications et d'additions [6]. Toute fois, elles manquent de précision. De variétés de techniques temporelles sont décrites dans la littérature. Parmi les techniques de base on peut citer : la Fonction d'AutoCorrélation (FAC) et ses versions modifiées [6], la Fonction de différence d'AMDF (Average Magnitude Difference Function) et ses variantes [7], la Fonction de réduction de donnée, DARD (DAta ReDuction method) [8] et la Fonction du calcul parallèle, PPROC (Parallel PRoCessing method) [9].

Les méthodes spectrales sont définies comme étant celle qui permet d'obtenir une F_0 en traitant le spectre de la parole directement. Parmi ces techniques, on peut citer : la technique Cepstrale (CEP) [10], le Produit Harmonique Spectral (HPS), et l'intercorrélation avec le Peigne Spectrale (PS) [2].

Les méthodes hybrides, visent à combiner différentes approches pour augmenter les performances globales du système d'extraction. Elles appliquent différents analyseurs simultanément sur le signal et combinent les différents estimateurs [5].

4 Description des techniques

Dans la plupart des algorithmes d'extraction de la F_0, trois phases essentielles durant le traitement s'implique : le prétraitement, le traitement et le pos traitement.

La phase de prétraitement est réservée à la préparation du signal issue d'un microphone. Elle consiste à choisir la durée des trames d'analyse et du recouvrement afin de moins compromettre la condition de stationnarité exigée par les algorithmes de traitement et l'effet de bord lié aux fenêtres de pondération appliquées.

La durée de la trame est généralement choisie entre 20 et 50ms avec un recouvrement de 30 à 50%, pour assurer la présence d'au moins une période du Fondamental [1]. Nous trouvons souvent d'autres techniques permettant d'améliorer la rapidité d'extraction tel que le filtrage, la décimation et les techniques de transformation non linéaire dites Clippage. La phase de traitement est réservée à l'extraction de la F_0 et dépend donc de l'algorithme utilisé.

La phase de post traitement à pour but de diminuer les erreurs qui peuvent être de plusieurs types. Ces erreurs vont être détaillées au cinquième paragraphe. On présente dans ce paragraphe les techniques choisies pour une éventuelle évaluation des performances.

4.1 La fonction d'autocorrélation basée sur le clippage central

Elle a été à l'origine proposée par L. Rabiner [3]. L'appellation Anglophone de cette technique est dite MACC (Modified Autocorrelation with Center Clipping) (Fig.1).

Le processus commence avec un filtre passe-bas, dont le but est d'atténuer l'influence des fréquences autres que la F_0. Le filtre coupe à 900 Hz, du fait qu'une valeur de pitch est comprise entre 70 et 600 Hz. La deuxième phase de traitement est la segmentation du signal vocal à des trames de 30 ms pour assurer la stationnarité du signal. La troisième phase est le calcul du seuil de clippage (C_L) pour chaque trame

d'analyse par la recherche les deux pics maximums dans la première (P_1) et la troisième (P_3) portion de 10 ms et de prendre le minimum de ces deux valeurs. Ce minimum est multiplié par la suite avec un niveau de clippage k. C'est un paramètre très important qu'il faut l'optimiser avec soin. On prend en général des valeurs variant entre 30% et 80% de l'amplitude de l'échantillon maximal de la trame [11]. La fonction de clippage qui est implémentée ici, est le clippage central avec compression [5]:

$$y(n) = clc[x(n)] = \begin{cases} x(n) - C_L & x(n) \geq C_L \\ 0 & |x(n)| < C_L \\ x(n) + C_L & x(n) \leq -C_L \end{cases} \quad (\mathbf{1})$$

La quatrième phase de traitement est le calcul de la FAC normalisée et la recherche du pic maximum (P) et son indice (ind. P) dans la gamme d'existante de la F_0 qui nous permettra par la suite de calculer la valeur du Fondamental.

La dernière phase de cet algorithme consiste à choisir un seuil de décision du voisement (V_0) en fonction du pic calculé. Si le pic maximum de chaque trame obtenue lors de la phase précédente dépasse le seuil de voisement, la trame est classifiée comme Voisée, sinon elle est classifies Non-Voisée. Dans le cas du silence, la détection se fait grâce à l'énergie à courte terme suivant un seuil bien défini. Si la valeur de l'énergie dans chaque trame ne dépasse pas ce seuil, la trame est considérée comme silence [3].

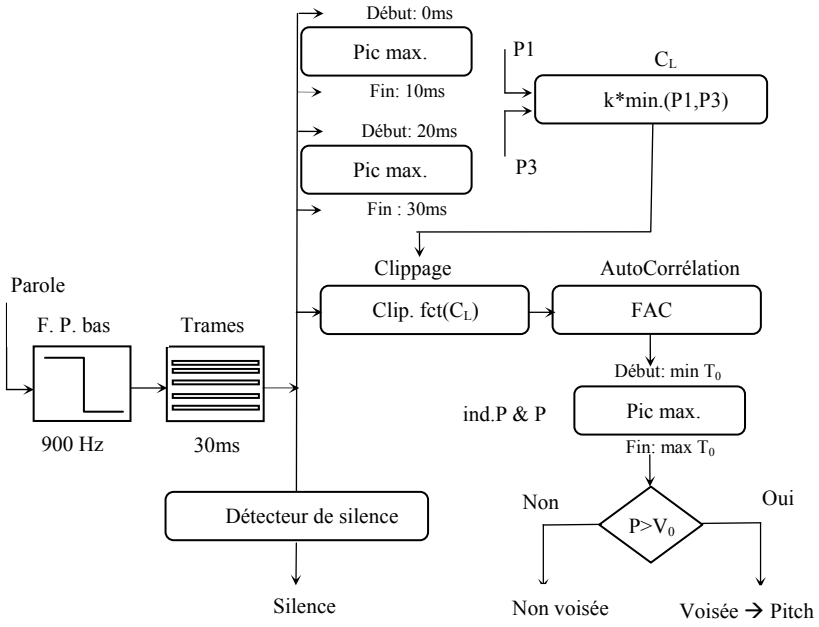

Fig. 1: Schéma bloc du détecteur de pitch par la MACC

4.2 La fonction d'AMDF basée sur le Clippage

Plusieurs versions d'AMDF existent pour la détection de la F_0 [7]. On a limité notre étude sur la C-AMDF (Clipping -Average Magnitude Diffrence Function). En premier lieu, le signal vocal est filtré par un filtre passe bas de type Butterworth à une Fréquence de coupure Fc de 900 kHz. Ensuite, segmenté en trames de 30 ms.

L'opération de clippage consiste à appliquer sur les fenêtres à court termes résultantes une transformation non linéaire définit par le clippage central donné par l'équation (1). Un seuil de clippage C_L doit être calculé pour chaque trame. Dans notre application, le seuil de clippage est choisi égal à 30% de l'amplitude du pic maximal de la trame en cour de traitement. L'AMDF est calculée sur le signal clippé pour chaque trame d'analyse. La valeur de la F_0 est déterminée avec la localisation de la vallée minimale entre 70 Hz et 600 Hz. En fin, la décision du voisement-silence s'effectue avec le calcul du Taux de Passage par Zéros (TPZ) et de l'énergie (Fig.2).

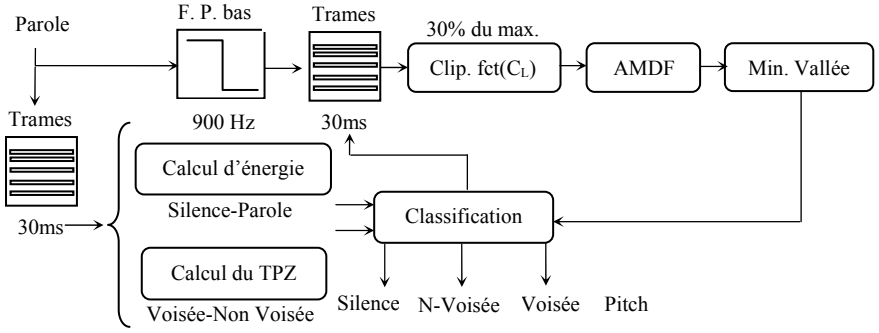

Fig. 2: Schéma bloc proposé du détecteur de pitch par la C-AMDF

4.3 La technique Cepstrale

L'estimation de la période de pitch peut être faite sur le Cepstre réel. La Figure 3 représente la description du détecteur de pitch par la méthode Cepstrale [3]. Chaque segment de 51.2ms est pondéré par une fenêtre de type hamming.

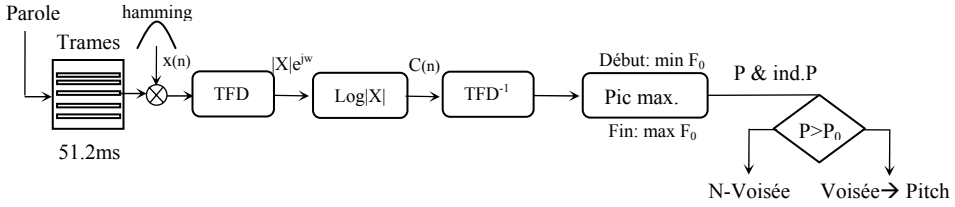

Fig. 3: Schéma bloc proposé du détecteur de pitch par la méthode Cepstrale

Le principe de la procédure de calcul de pitch fondé sur le Cepstre est plutôt simple. On recherche dans le Cepstre un pic dans la région autour de la période du pitch (P). Si le pic est supérieur à un seuil fixé (P_0), le segment de parole en entrée est probablement Voisé, et la position autour du pic est la zone dans laquelle on peut estimer le pitch. Si le pic n'est pas supérieur au seuil, il est alors probable que le segment de parole en entrée est non Voisé [12].

4.4 Le Produit Harmonique Spectral

La méthode HPS, pour Harmonic Product Spectrum (HPS), a été publiée pour la première fois par R. Noll en 1970 [13]. Cette méthode est basée sur le principe de la compression des raies spectrales (Fig.4). Pour chaque trame stationnaire du signal vocal x(n) (d'une durée de 30ms), le logarithme de sa densité spectrale de puissance est calculé le long de l'axe des fréquences sur des facteurs entiers. La valeur logarithmique de l'HPS est obtenue par l'addition du spectre original et ses versions compressées (décimées) [2]:

$$\mathrm{HPS}(m) = \sum_{r=1}^{R} \log\left|x(rm)\right|^2 \qquad (\mathbf{2})$$

R représente le nombre total des spectres impliqués dans le calcul et X(k) la Transformée de Fourier Discrète (TFD) de x(n). Pour obtenir le Produit Spectral d'Harmonique, l'exponentiel de la fonction doit être pris.

Le choix de la constante R joue un rôle principal sur la précision du détecteur. De nombreux travaux de recherche étaient basés fondamentalement sur un facteur de décimation R=5. C'est un choix qui offre une meilleure estimation de la F_0 pour une Fréquence d'échantillonnage (F_e) de 16kHz [7,14].

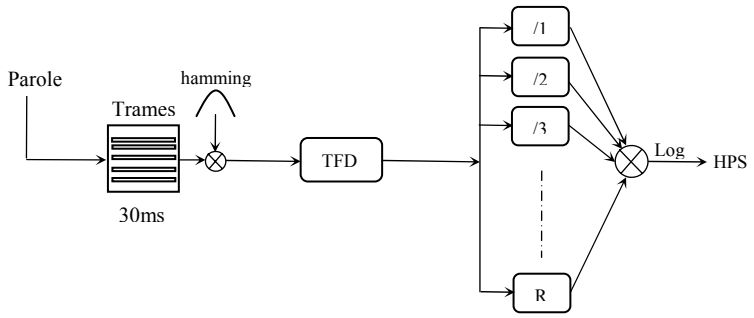

Fig. 4: Schéma bloc du détecteur de pitch par l'HPS

Quand les spectres compressés sont ajoutés, les harmoniques présentes dans le signal de parole s'ajoutent de manière constructive, puisqu'ils sont multiples de la F_0.Les composantes fréquentielles du bruit et des sons non Voisés, ne montrent pas le même rapport, par conséquent seront noyées par l'opération de la somme [2]. Pour la classification des sons, on a suivi la même procédure de la C-AMDF basé sur le TPZ et l'énergie à court terme de chaque trame.

5 Signaux de tests et F_0 de référence

Les signaux de tests utilisés dans notre étude pour l'évaluation des algorithmes de détection de la F_0 sont classifiés en deux groupes : des sons Voisés purs et des sons mixtes (Voisement et silence). C'est un corpus qui englobe les catégories des sons de l'AS, à savoir, les voyelles (orales et nasales) et les consonnes (plosives, fricatives, nasales, liquides, vibrantes, affriquées et semi voyelles) [16]. Pour la première catégorie, et du fait que le voisement est une caractéristique importante dans les algorithmes de détection de F_0, le choix est porté sur le phonème [a], une voyelle pure, prononcé par trois locuteurs de différents âges et sexes (masculin âgé de 25ans, féminin de 20ans et enfantin de 5ans). Les sons sont enregistrés pendant une durée de 2.5 secondes et échantillonnés à une Fe de 16 kHz.

Pour la deuxième catégorie, et afin évaluer les performances des algorithmes à effectuer des classifications automatiques, on a utilisé deux différentes phrases prononcées en AS. La première phrase est prononcée par un locuteur masculin, elle est caractérisée par une durée de 6s. La deuxième phrase est prononcée par un locuteur féminin et est d'une durée de 9s. Les deux phrases sont échantillonnées à 16kHz. Pour les quatre détecteurs, une optimisation des paramètres de chaque algorithme par des tests pratiques est faite pour une bonne estimation de la F_0.

1. Phrase 1 : « وهي في موقع عند أقصر مسافة بين الدجلة و الفرات »
2. Phrase 2 : « والعامية من ناحية أ خرى ليست واحدة بل لهجات متعددة»

Les valeurs de référence de la $F_{0réel}$ dite, Fréquence Fondamentale pour une analyse standard [3], sont mesurées manuellement pour chaque trame des signaux de tests choisis. Les zones Non Voisées et silences correspondent à une F_0 nulle. D'une manière globale, on ne peut pas dire qu'on est doté d'une bonne base de données des sons, mais d'un corpus modeste qui nous a permis de comparer les performances des techniques choisis sur des sons prononcés en AS. Le corpus est d'une durée limitée du fait que les valeurs de $F_{0réel}$ sont mesurées manuellement. C'est une opération difficile mais valable pour bonne évaluation des paramètres d'erreurs. C'est la même approche utilisée dans [7].

6 Paramètres d'erreurs

Plusieurs paramètres d'erreur d'estimation de la F_0 peuvent être employés pour évaluer la qualité d'un algorithme d'extraction. On s'est limité dans notre étude aux paramètres principaux. Soit $F_{0reel}(m)$ et $F_{0j}(m)$ respectivement les valeurs réelles (analyse standard) et estimées de la F_0 de chaque signal de test. Soit m l'indice de trame qui varie selon la taille du signal d'entrée et j un indice qui représente la technique d'estimation de la F_0 variant de 1 à 4, respectivement pour la MACC, C-AMDF, CEP et HPS. Les paramètres d'erreur sont élaborés suivant quatre possibilités :

1. $F_{0reel}(m) = 0$ et $F_{0j}(m)=0$, dans ce cas, l'analyse standard et le j^{eme} détecteur du pitch classifient la m^{ieme} trame Non Voisée. Dans ce cas aucune erreur de calcul ne résulte.

2. $F_{0reel}(m) = 0$ et $F_{0j}(m) \neq 0$, l'analyse standard classifie la m^{ieme} trame Non Voisée, par contre le détecteur du pitch la classifie Voisée. Dans ce cas une erreur Non Voisée-Voisée (NV_V) est apparue. Ce type d'erreur est déterminé par la relation suivante [15] :

$$NV\_V = \frac{\text{Taille}(F_{0j} \neq 0 \,\&\, F_{0réel} = 0)}{F_{0réel} = 0} \quad (\mathbf{3})$$

Où « Taille $(F_{0reel}=0)$ » représente le nombre de trames où $F_{0réel}$ est nulle et « $\text{Taille}(F_{0j} \neq 0 \,\&\, F_{0réel} = 0)$ » représente le nombre de trames où à la fois $F_{0réel}$ est nulle et F_{0j} est différent de zéro.

3. $F_{0reel}(m) \neq 0$ et $F_{0j}(m)=0$, l'analyse standard classifie la m^{ieme} trame Voisée, par contre le détecteur du pitch la classifie Non Voisée. Dans ce cas une erreur Voisée-Non Voisées (V_NV) est apparue. Ce type d'erreur est déterminé par la relation suivante [15]:

$$V\_NV = \frac{\text{Taille}(F_{0j} = 0 \,\&\, F_{0réel} \neq 0)}{F_{0réel} \neq 0} \quad (\mathbf{4})$$

Où «Taille $F_{0réel} \neq 0$ » représente le nombre de trames où $F_{0réel}$ est différent de zéro et « $\text{Taille}(F_{0j} = 0 \,\&\, F_{0réel} \neq 0)$ » représente le nombre de trames où à la fois $F_{0réel}$ est différent de zéro et F_{0j} est nulle.

4. $F_{0reel}(m)= P1 \neq 0$ et $F_{0j}(m)= P2 \neq 0$, la m^{ieme} trame est classifiée Voisée dans les deux cas. Trois types d'erreur dépendent des valeurs de P_1 et P_2. L'erreur du Voisement e(m) est définit comme suite [3] :

$$e(m) = |P_2 - P_1| \quad \mathbf{(5)}$$

4.1. Si $e(m) \geq 16$ échantillons, (plus de 1ms d'erreur d'estimation de la T_0 pour une Fe de 16kHz) [3,14]. L'erreur d'analyse est considérée comme une erreur grossière.

Pour de tels cas, le détecteur de pitch a nettement échoué en estimant la F_0. Les causes possibles de ces erreurs sont le doublement ou le triplement de la F_0 ;

4.2. Si $e(m) < 16$ échantillons, l'erreur d'analyse est classifiée comme une erreur fine. Pour de tels cas le détecteur de pitch a estimé la F_0 d'une manière suffisamment exacte ;

4.3. Si $e(m) = 0$ pas d'erreur d'analyse commise car les deux valeurs P_1 et P_2 sont égaux. C'est un cas particulier de 4.2 nommé PESEC (Paramètre d'Extraction Sans Erreur Commise). Il caractérise un pourcentage dérivé des erreurs fines dont l'erreur d'estimation est exactement nulle.

7 Performances des algorithmes basées sur la somme des erreurs

D'après les travaux de L. Rabiner [3], l'évaluation globale des performances des estimateurs s'effectue par une sommation des paramètres d'erreur de la base de

données utilisée. Dans notre étude, la sommation des erreurs est établie pour tous les sons choisis (Tableau 1).

Tableau 1: Paramètres d'erreurs globales (en pourcentage)

Voix	Erreurs	MACC	CAMDF	CEP	HPS	somme
Phrase 1	Fines	80.16	78.22	93.54	42.69	294.61
	Grossières	19.84	21.78	6.46	57.31	105.39
	PESEC	5.52	6.30	6.03	0.69	18.54
	V_NV	4.8	0.8	53.03	0.8	59.43
	NV_V	51.94	54.54	3.85	54.54	164.87
Phrase 2	Fines	76.91	89.69	97.46	73.15	337.21
	Grossières	23.09	10.31	2.54	26.85	62.79
	PESEC	4.14	7.60	3.70	0.59	16.03
	V_NV	19.16	1.19	15.95	1.19	37.49
	NV_V	24.81	38.68	23.80	38.68	125.97
Phonème [a] masculin	Fines	100	100	100	92.85	392.85
	Grossières	0	0	0	7.15	7.15
	PESEC	3.57	1.2	5.96	7.73	18.46
Phonème [a] Féminin	Fines	100	100	100	95.35	395.35
	Grossières	0	0	0	4.65	4.65
	PESEC	2.33	2.33	3.49	0.01	8.16
Phonème [a] Enfantin	Fines	100	100	100	100	400
	Grossières	0	0	0	0	0
	PESEC	1.17	1.17	1.17	16.27	19.78
Somme	**Fines**	457.07	467.91	491	404.04	**1820**
	Grossières	42.93	32.09	9	95.96	**179.98**
	PESEC	16.73	18.60	20.35	25.29	**80.97**
	V_NV	23.96	1.99	68.98	1.99	**96.92**
	NV_V	76.75	93.22	27.65	93.22	**290.84**

La lecture du Tableau 1 nous permettra d'obtenir une évaluation détaillée des techniques utilisées en fonction des signaux de test en entrée. Les erreurs globales obtenues par la sommation des erreurs calculées (horizontales et verticales) nous permettent de faire une appréciation sur les techniques utilisées (sommations horizontales). Elles nous permettent aussi de faire une mesure d'exactitude d'estimation de la F_0 de chaque signal de test avec l'utilisation des quatre détecteurs (sommations horizontales). D'après les erreurs globales d'estimation de la F_0 obtenues par les tous les estimateurs (sommation verticale), on peut dire que la technique Cepstrale offre la meilleure estimation avec moins d'erreurs grossières commises (doublement et triplement du pitch) (Fig. 5, signal d'entrée en bleu, $F_{réel}$ en noire et F_0 estimée en rouge). Elle présente aussi un bon PESEC par rapport aux deux techniques temporelles utilisées. Cependant, elle présente un taux élevé d'erreurs

V_NV. Les deux techniques temporelles respectivement la MACC et C-AMDF sont classées en second et troisième position. Elles présentent des taux d'erreurs globales et des PESECs comparables. Néanmoins, elles présentent un taux d'erreurs NV_V élevé.

On remarque aussi que la technique de classification utilisée par la C-AMDF ainsi que l'HPS présente le meilleur score d'erreurs V_NV (1.99). C'est une caractéristique importante dans ce type de traitement.

En dernière position vienne la technique basée sur l'HPS, possédant le mauvais score d'erreurs grossières. Toutefois, elle présente le meilleur PESEC qui est une caractéristique intéressante offerte par cette technique. On remarque que dans le cas des sons voisés purs (Phonèmes [a]), l'estimation de F_0 est faite d'une manière suffisamment exacte par toutes les techniques (sommation horizontales). C'est un résultat attendu du fait que les voyelles de l'AS sont caractérisées par un taux de voisement important. En dernier lieu, on peut dire que notre système d'évaluation répond à une certaine hypothèse d'ergodicité du fait que la sommation des erreurs fines, grossières, voisement et de PESECs verticales et toujours égales à celles horizontales.

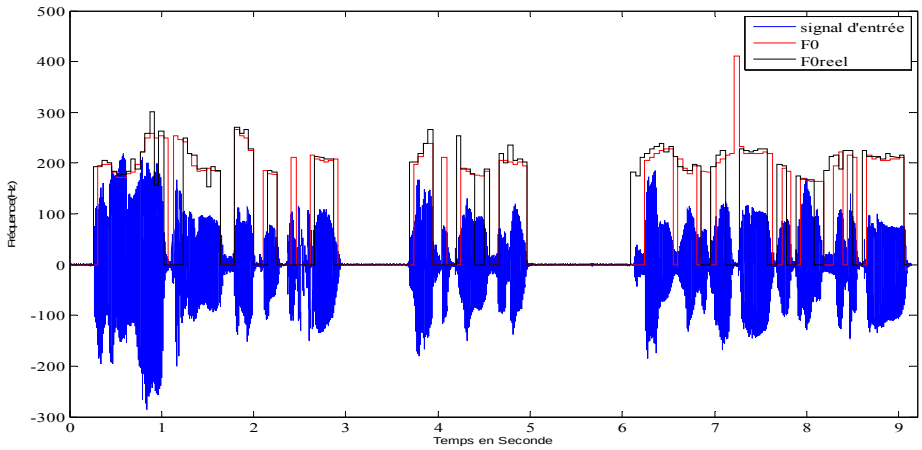

Fig. 4 : Comparaison entre la F_{0reel} et la F_0 estimée de la phrase 2 par la CEP

8 Conclusion

Nous avons présenté dans cet article une évaluation des performances de plusieurs techniques d'estimation de la fréquence Fondamentale du signal vocal en se basant sur des sons prononcés en Arabe Standard par des locuteurs de différents sexes et âges. On a déduit que la CEP a donnée une meilleure estimation de la F_0 pour chaque locuteur par rapport aux autres techniques utilisées. Cependant, l'estimation de la F_0

par l'HPS a donnée la mauvaise estimation. C'est un résultat logique du fait que cette technique est essentiellement utilisée pour des sons musicaux plutôt que de parole.

La méthode de segmentation par fenêtre fixe utilisée par la C-AMDF et l'HPS présente des bonnes performances à détecter les régions de transition V_NV en la comparant a la méthode de segmentation par seuillage utilisée par la MACC et la technique Cepstrale. Le facteur PESEC introduit au niveau de cette étude nous a permis de découvrir que les estimations fines faites par l'HPS sont plus exacte malgré le taux des erreurs grossières marqué. Le corpus modeste employé nous a permis de faire un test d'évaluation pratique des performances des techniques implémentées. La durée du corpus est réduite du fait que les valeurs de l'analyse standard (F_0 réel) sont prises manuellement. L'élaboration d'une large base de données des sons spécifiques sur plusieurs conditions d'enregistrement (Téléphoniques, proche ou loin du microphone et la qualité du microphone utilisé) nous permet de mieux évaluer les performances des techniques employées sur des sons en Arabe Standard. Cette évaluation va nous aider au développement des nouvelles techniques hybrides d'estimations en exploitant les avantages des techniques étudiées.

Références

1. Boite, R.:Traitement Automatique de la Parole. Edition Masson, France (1989).
2. Flego, F.: Fundamental Frequency Estimation Techniques for Multi Microphone Speech Input. Phd Dissertation, University of Trento, USA (2006).
3. Rabiner, L.R.: A Comparative Performance Study of Several Pitch Detection Algorithms. IEEE Trans.Acoust., Speech, And Signal Processing, Vol. ASSP-24, No.5, October (1976).
4. Bimbot, F., Magrin-Chagnolleau, I, Mathan, L.: Second-order statistical measures for text-independent speaker identification. Speech Communication, Volume 17, Number 1, pp. 177-192(16), Elsevier (1995).
5. Hess, W.: Pitch Determination of Speech Signals: Algorithms and Devices. Edition Springer-Verlag, Berlin (1983).
6. Dubnowski, J., Schafer, R. W., Rabiner, L. R.: Real-Time Digital Hardware Pitch Detector. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 2-8 (1976).
7. Yu-Min Zeng and al.: Modified AMDF Pitch Detection Algorithm. Proceedings of the Second International Conference on Machine Learning and Cybernetics Wan, 2-5 (2003).
8. Miller, N. J.: Pitch Detection by Data Reduction. IEEE Tranr, Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 72-79 (1975).
9. Rosenberg, A. E., Sambur, M. R.: New Techniques for Automatic Speaker Verification. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 169-176 (1975).
10. Schafer, R. W., Rabiner, L. R.: System for Automatic Formant Analysis of Voiced Speech. J. Acoust. Soc. Amer., vol. 47, pp. 634-648 (1970).
11. Jean Laroche: Cours sur le Traitement des Signaux Audio-Fréquences. Département du Signal, Groupe Acoustique-Telecom Paris (1995).
12. Vũ Minh Quang: Exploitation de la Prosodie pour la Segmentation et l'Analyse Automatique des Signaux de Parole. Thèse de Doctorat, Institut Polytechnique de Hanoi, France (2007).
13. Van Doremalen: Procédé d'Extraction de la Fréquence Fondamentale d'un Signal Vocal. Office des Brevet Europèen, 0 821 345 A1 EP, Paris (1998).

14. Huin Ding, Bo Qian: A Method Combining LPC Based Cepstrum and Harmonic Product Spectrum for Pitch Detection. Proceeding of the IEEE International Conference on Intellegent Information Hiding and Multimedia Signal Processing,IIH-MSP 06,USA (2006).
15. Kavita Kasi: Yet another Algorithm for Pitch Tracking. Master's thesis, Old Dominion University, UK (2002).
16. Droua-Hamdani, G.: Prédiction de la Durée Segmentale des Phonèmes de l'Arabe Standard. Mémoire de Magister, CRSTDLA, Alger, Algérie (2004).