Estimation du pitch et décision de voisement par compression spectrale de l'autocorrélation du produit multi-échelle

Mohamed Anouar Ben Messaoud, Aïcha Bouzid et Noureddine Ellouze
Laboratoire signal, image et Technologies de l'Information, ENIT Le Belvédère, B.P.37, 1002, Tunis
anouar.benmessaoud@yahoo.fr,
bouzidacha@yahoo.fr,n.ellouze@enit.rnu.tn

RESUME

Dans ce papier, nous proposons un algorithme d'estimation de la fréquence fondamentale et de décision de voisement à partir des signaux de parole. Notre approche est basée sur la décimation du spectrale numérique de la fonction d'autocorrélation du produit multi-échelle (APM) du signal de parole. Le produit multiéchelle est le produit des coefficients de la transformée en ondelettes du signal de parole calculées à différentes échelles successives. Le pitch est estimé par la multiplication des copies comprimées du spectre original de l'APM. Le signal obtenu permet d'opérer la décision de voisement et l'estimation de pitch. Nous présentons une méthodologie d'évaluation qui associe la décision de voisement dans la procédure d'estimation du pitch et présente une étude comparative de la performance des algorithmes d'estimation du pitch qui montre l'impact de la décision de voisement sur les résultats d'estimation de F_0.

ABSTRACT

Pitch estimation and voiced decision by spectral autocorrelation compression of multi-scale product

In this work, we propose an algorithm for pitch estimation and voicing detection in clean and noisy speech signal. Our approach is based on the spectral compression (CS) of the autocorrelation of the multi-scale product (APM). The APM consists of making the autocorrelation of the speech wavelet transform coefficients product at three successive dyadic scales. We estimate the pitch for each frame based on the product of compressed copies of the original spectrum of APM. To make the voiced/unvoiced decision, we use the estimated F0 value. In addition, we present a Gross Pitch Classification Error methodology which add Gross Pitch Error and Voicing Classification Error to measure the robustness of pitch determination algorithms and to show the impact of the voiced/unvoiced decision on the results obtained by any pitch determination algorithms.

MOTS-CLES : Autocorrélation du produit multi-échelle, compression spectrale, estimation du pitch, décision de voisement.
KEYWORDS : Autocorrelation multi-scale product, spectral compression, pitch estimation, voicing decision.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 201–208,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les caractéristiques du signal vocal englobent entre autres des paramètres de source liés aux vibrations des cordes vocales comme le voisement et la fréquence fondamentale. Le pitch est un paramètre fondamental dans la production, l'analyse et la perception de la parole. Ce paramètre est un révélateur principal de l'information phonétique, lexicale, syntaxique et émotionnelle. Il entre dans la mise au point de diverses techniques avancées d'analyse et d'interprétation du signal de parole et devient en ce sens un élément essentiel dans la mise en œuvre des applications en traitement automatique de la parole (Saito, 1992).

La problématique liée au pitch et sa complexité apparaît dans la multitude d'algorithmes de détermination du pitch (ADPs). Ces algorithmes ne présentent pas les mêmes performances pour tous les types de voix et dans toutes les conditions (Hermes, 1993). Certains algorithmes ont fait leurs preuves sur des signaux de parole. D'autres applications exigent plus de précision. En effet, nous discernons une multitude d'algorithmes pour l'estimation du pitch (Gerhard, 2003). Les algorithmes les plus récents proposent de nouvelles approches ou essaient d'améliorer des méthodes existantes. Dans ce papier, nous présentons une nouvelle approche de détermination du voisement et d'estimation du pitch basée sur la compression spectrale de l'autocorrélation du produit multi-échelle du signal de parole, la méthode est comparée à d'autres méthodes pour évaluer ses performances et sa robustesse.

Les ADPs ne sont performants que si l'évaluation de la fréquence fondamentale est liée à une décision de voisement fiable. Ainsi, l'évaluation d'erreur grossière (GPE) implique l'évaluation d'erreur de classification (CE). Les performances de ces divers algorithmes seront validées selon cette relation entre GPE et CE.

Le papier est présenté comme suit. Après l'introduction, nous présentons notre approche d'estimation de la fréquence fondamentale et la décision de voisement. Dans la Section 3, nous présentons la métrique employée pour l'évaluation des performances et proposons une méthodologie d'évaluation qui prend en compte les deux paramètres GPE et CE. Dans la section 4, nous présentons l'influence de la décision de voisement sur les résultats de l'évaluation du pitch par une comparaison avec d'autres ADPs.

2 Algorithme proposé

L'Algorithme proposé est basé sur l'analyse de la compression spectrale de la fonction d'autocorrélation du produit multi-échelle (CSAPM). Cette approche peut être répertoriée dans la classe de la détermination simultanée de voisement et du pitch. Notre approche vérifie la présence d'une condition suffisante pour décider qu'une trame est voisée, la fréquence F_0 estimée est employée ensuite pour classifier la trame en voisée ou non voisée. Des critères supplémentaires sont ajoutés pour conclure au non voisement.

2.1 Estimation du pitch

Dans cette section, nous présentons l'approche CSAPM pour l'estimation de la fréquence fondamentale. Cette estimation est précédée par le calcul de l'énergie de la trame de la

parole, qui est considérée comme condition suffisante préalable pour décider que la trame est voisée ou non voisée. La méthode de compression spectrale de l'auto-corrélation du produit multi-échelle par (CSAPM) résumée en trois étapes est schématisée par la figure 1.

La première étape de l'algorithme consiste à calculer les transformées en ondelettes du signal de parole à trois échelles dyadiques successives, puis opérer la multiplication des coefficients pour obtenir le signal produit *p*. Le calcul du produit des coefficients de la transformée en ondelettes du signal de parole aux échelles ½, 1 et 2 avec l'ondelette spline quadratique comme ondelette mère de support $T = 0.8$ *ms*, permet d'obtenir un signal simplifié tout en gardant les propriétés de périodicité.

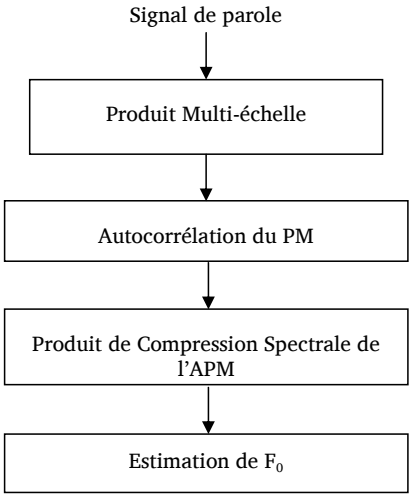

FIGURE 1 – Schéma block de l'algorithme d'estimation du pitch par la méthode CSAPM.

$$p_w[n, i] = p[n]\, w[n - i\Delta n] \tag{1}$$

Le PM $p(n, s_1, s_2, s_3)$ est pondéré par une fenêtre glissante $w[n]$:

i est l'indice de la fenêtre et Δn est l'intervalle de recouvrement.

La seconde étape consiste à opérer l'autocorrélation du produit multi-échelle (APM).

L'autocorrélation du signal *p* est calculée selon l'équation suivante :

$$A_i(k) = \sum_{n=0}^{N-1} p_{wi}(n)\, p_{wi}(n+k) \tag{2}$$

La troisième étape consiste à compresser le spectre de l'APM le long de l'axe fréquentiel selon différents facteurs de compression (R = 1, 2, 3, 4), puis de multiplier le spectre original à ses versions compressées. Ainsi les harmoniques s'alignent et renforcent la fréquence fondamentale.

Cette étape s'exprime par l'équation suivante :

$$C_i(k) = \prod_{r=1}^{R-1} FFT\left(A_i(r*k)\right) \qquad (3)$$

R représente le nombre total des spectres compressés dans le calcul. Le choix de ce paramètre joue un rôle principal sur la précision de l'estimateur. Le facteur de compression est choisi selon les résultats empiriques.

Pour montrer, la robustesse de notre approche, nous traitons le cas des extrémités des zones voisées et des zones voisées fortement bruitées.

La figure 2 montre le signal de parole au début d'une zone voisée prononcée par une femme suivi de son PM, son APM et enfin par son CSAPM. La compression spectrale d'APM fait ressortir la fréquence fondamentale.

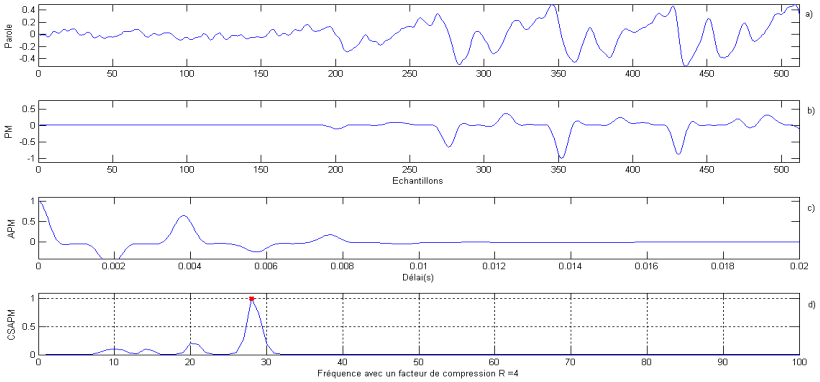

FIGURE 2 – CSAPM au début d'une zone voisée prononcé par une femme a) début d'une voyelle b) son PM c) son APM d) son CSAPM.

La figure 3 montre le signal de parole voisé corrompue par un bruit blanc gaussien à un RSB de -5 dB suivi de son PM, APM et CSAPM. La compression spectrale d'APM donne une raie claire sans bruit qui représente la fréquence fondamentale. On observe la capacité de CSAPM de produire des raies nettes.

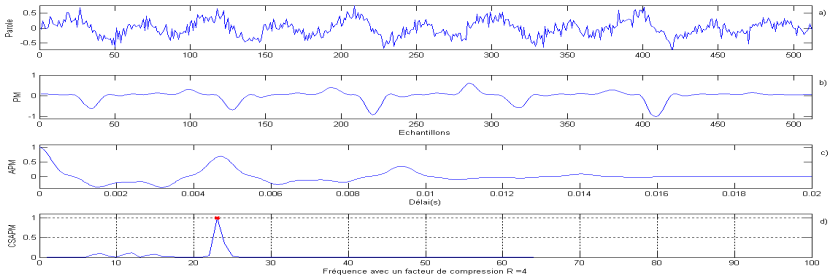

FIGURE 3 – CSAPM d'une voyelle prononcée par une femme corrompue par un bruit blanc gaussien à un RSB de - 5 dB. a) Signal de parole voisé bruité b) son PM c) son APM d) son CSAPM.

2.2 Décision de voisement

En utilisant la valeur de F_0 déjà estimée, on sélectionne deux périodes au milieu de la trame. Si la corrélation entre les deux segments dépasse un seuil S1 et le taux de passage par zéro dépasse un seuil S2 alors la trame est considérée comme voisée. Dans le cas contraire, la trame est considérée comme non voisée et la valeur de la fréquence F0 s'annule. Cette stratégie assure également la détection des signaux semi-voisés.

3 Méthodologies d'évaluation

3.1 Méthodologie classique

Afin d'évaluer les performances d'une approche de détection de voisement proposée, nous déterminons:

- Le taux d'erreurs voisé/non voisé (V/NV) qui constitue les zones voisées qui sont classées comme non voisées ; il s'agit de mesures manquées.
- Le taux d'erreurs non voisé/voisé (NV/V) qui constitue les zones non voisées qui sont classées comme voisées; il s'agit de fausses alarmes.

Pour l'estimation du pitch, nous comparons la mesure du pitch de référence opérée sur le signal électroglottographique (EGG) à celle donnée par l'approche proposée sur le signal de parole. La valeur absolue de la différence entre les valeurs des fréquences fondamentales de référence et les fréquences fondamentales estimées constitue l'erreur absolue. Lorsque l'erreur est inférieure ou égale à 20% de la valeur référence, elle est comptée comme une erreur fine. Les erreurs dépassant 20% sont comptées comme grossières.

3.2 Influence de la décision de voisement sur l'estimation du pitch

La décision de voisement marque si la trame analysée possède une fréquence fondamentale ou non ce qui montre que cette décision influe les résultats de l'évaluation

de l'estimation du pitch. Ainsi, la mesure de voisement et de la fréquence fondamentale sont deux concepts fortement liés pour assurer une comparaison significative entre les ADPs. En effet, la décision de voisement dépend souvent de l'estimation du pitch et inversement (Ghio, 2007), (Signol, 2008).

L'influence de la décision de voisement sur l'estimation de F_0 se voit surtout au niveau des extrémités des zones voisées. Un ADP qui considère ces zones commet plus d'erreurs d'estimation de F_0 que s'il ne les considère pas.

3.3 Méthodologie proposée

L'influence de la décision de voisement sur les résultats d'estimation du pitch a donné lieu à une méthodologie d'évaluation associant la décision de voisement dans la procédure d'estimation proposée. Cette méthodologie additionne les paramètres d'erreur d'estimation de F_0 aux paramètres de voisement pour évaluer l'algorithme et faire une étude comparative de la performance des ADPs. Nous proposons alors de calculer le taux suivant :

$$GPCE = \frac{(Nbr\ de\ trames\ déclarées\ voisées\ par\ F_0\ référence\ \&\ par\ F_0 estimé)}{(Nbr\ total\ de\ trames)} * GPE + CE \quad (4)$$

Avec

$$GPE = \left| F0_{référence} - F0_{estimé} \right| \Big/ F0_{référence} > 0.2 \quad (5)$$

Et

$$CE = \frac{\left(Nbr\ Zone\ V/NV + Nbr\ Zone\ NV/V\right)}{\left(Nbr\ total\ de\ trames\right)} * 100\% \quad (6)$$

L'équation (4) est obtenue par la multiplication du GPE au nombre de trame considérée voisées par la F_0 référence et la F_0 estimée, en rajoutant le terme de l'erreur de classification. Nous pouvons alors comparer les différents ADPs de façon similaire.

4 Evaluation des Algorithmes de décision de voisement et de détermination de F0

4.1 Conditions d'évaluation

Pour évaluer et comparer des ADPs de manière équitable, il faut se mettre dans les mêmes conditions de travail à savoir la base de sons utilisée, considérer les zones de fin de voisement et l'intervalle de variation de F_0.

4.2 Etude comparative avec les méthodes existantes

L'évaluation est opérée sur la base de son de l'Université de Keele. Il s'agit de dix locuteurs ayant l'anglais comme langue maternelle et prononçant le texte « The North Wind Story ». Ces locuteurs sont 5 hommes âgés de 21 à 60 et 5 femmes âgées de 20 à

37 ans (Plante, 1995). La base de Keele a été développée dans l'objectif d'évaluer les performances des algorithmes de détection de voisement et de détermination du pitch. Pour satisfaire cet objectif, le signal de parole et le signal EGG pris comme signal de référence, ont été enregistrés simultanément dans une pièce insonorisée. Les deux signaux sont par la suite échantillonnés à une fréquence de 20 kHz et codés sur 16 bits. Pour tous les algorithmes évalués, nous utilisons une fenêtre de longueur 25.6 ms avec une estimation de F_0 pour chaque 10 ms.

En appliquant la méthodologie d'évaluation proposée, le tableau 1 récapitule les performances de différents algorithmes en utilisant la base de Keele dans un environnement non bruité. La méthode SWIPE' présente le plus faible taux d'erreurs grossières de 0.62% alors que notre approche CSAPM a un taux GPE légèrement supérieur de 0.67%. Par contre, lorsque nous prenons en considération l'influence de voisement sur le pitch notre méthode décroche la meilleure performance avec le plus faible pourcentage GPCE qui est de 2.59 %.

Méthodes			
	GPE (%)	CE (%)	GPCE (%)
CSAPM	***0.67***	2.27	***2.59***
SPM (Ben Messaoud, 2010)	0.75	3.02	3.31
SWIPE' (Camacho, 2007)	***0.62***	3.92	***4.19***
YIN (De Cheveigne, 2002)	2.28	6.28	7.23

TABLE 1 – GPE, CE et GPCE sans bruit pour toute la base de Keele

Le tableau 2, présente la robustesse de notre approche CSAPM comparée à celles des algorithmes suivants : SPM (Ben Messaoud, 2010), SWIPE' (Camacho, 2007) et YIN (De Cheveigne, 2002) en présence de différents types de bruits (bruit blanc gaussien, bruit babble) à un RSB de -5 dB. Ces bruits sont extraits de la base NOISEX92 (Noisex92, 1992).

En effet pour le taux GPE, la méthode SWIPE' donne le meilleur résultat car elle ne considère que les trames fortement voisées. Alors que le taux GPCE montre bien que notre approche surpasse les autres méthodes dans les zones de faible voisement en présence de bruit.

	White Noise (RSB = -5 dB)			Babble Noise (RSB = -5 dB)		
	GPE (%)	CE (%)	GPCE (%)	GPE (%)	CE (%)	GPCE (%)
CSAPM	***1.12***	4.59	**5.06**	***1.73***	6.27	***6.94***
SPM	1.40	8.41	8.92	7.62	7.26	9.85
SWIPE'	***0.48***	43.44	***43.62***	***0.22***	51.67	***51.76***
YIN	5.33	7.32	9.50	6.14	5.38	8.08

TABLE 2 – GPE, CE et GPCE en présence de bruit en utilisant la base de Keele

5 Conclusion

Dans ce papier, nous avons proposé une méthode robuste d'estimation du pitch et de décision de voisement. La compression spectrale de l'autocorrélation du produit multi-échelle (CSAPM) procède au calcul du produit des coefficients de la transformée en ondelettes pour différentes échelles successives du signal de parole, puis au calcul de son autocorrélation. Ensuite, le spectre de l'APM subit un ensemble de compressions de facteurs entiers. Le produit des spectres compressés permet le rehaussement des harmoniques pour une meilleure estimation du pitch et une meilleure décision de voisement. L'évaluation proposée tient compte non seulement de l'erreur d'estimation de la fréquence F_0 mais aussi de la décision de voisement ce qui permettrait d'opérer une comparaison significative avec d'autres algorithmes. Les perspectives de ce travail concernent l'extension de l'approche proposée à l'estimation multi-pitch dans un contexte multi-locuteurs.

Références

BEN MESSAOUD, M.A., BOUZID, A. et ELLOUZE, N. (2011). Using multi-scale product spectrum for single and multi-pitch estimation. *In* (*IET Signal Processing Journal, Vol.5, N.3*), pages 344–355.

CAMACHO, A. (2007). SWIPE: A sawtooth waveform inspired pitch estimator for speech and music. Thèse de Doctorat,University of Florida, USA.

DE CHEVEIGNE, A. et KAWAHARA, H. (2002). YIN, a fundamental frequency estimator for speech and music. *In* (*J. Acoust. Soc. Amer., Vol.111, N. 4*), pages 1917–1930.

SIGNOL, F., Barras, C., LIENARD, J-S. (2008). Evaluation of the pitch estimation algorithms in the monopitch and multipitch cases, *In* ACOUSTICS 2008, Paris, France.

GERHARD, D. (2003). Pitch extraction and fundamental frequency: History and current techniques, Tech. Rep, Department of Computer Science, University of Regina, Canada.

GHIO, A. (2007). Evaluation acoustique. *In* (Auzou P.: Rolland V., Pinto S., Ozsancak C. Les dysarthries. Marseille: Solal. 2007), pages 236–247.

HERMES, D.J. éditeur Wiley, J. (1993). Pitch analysis, In visual representation of speech signals. *In* (Wiley, J., 2003), pages 1–25.

PLANTE, F. MEYER, G.F. et AINSWORTH, W.A. (1995). A Pitch extraction reference database. *In* EUROSPEECH 1995 (*European conference on speech communication and technology*), Madrid, Espagne.

SAITO, S. (1992). Speech science and technology. *In* (IOS Press, 1992), pages 481–484.

NOISEX92. (1992). Signal Processing Information Base (SPIB). The signal processing society. http://spib.rice.edu/spib/select_noise.html. [consulté le 15/4/2012].