Produit multiéchelle pour la détection des instants d'ouverture et de fermeture de la glotte sur le signal de parole

Aïcha Bouzid et Noureddine Ellouze

Laboratoire signal, image et reconnaissance de formes (LSTS-ENIT)
Le Belvédère, B. P. 37, 1002, Tunis
bouzidacha@yahoo.fr, N.Ellouze@enit.rnu.tn,

ABSTRACT

This paper deals with robust singularity detection in speech signal using multiscale product method. These singularities correspond to opening and closure instants of the glottis (GOIs and GCIs). Multiscale product method consists of computing the products of wavelet transform coefficients of the speech signal at appropriate adjacent scales. As wavelet modulus maxima are a tool for signal edge detection, first derivative of a Gaussian function, is used for detecting speech signal discontinuities. Speech Multiscale products enhance edge detection. The proposed method is evaluated comparing to the EGG signal references using the Keele University database. This method gives excellent results concerning GOI and GCI detection from speech signal.

1. INTRODUCTION

Les vibrations quasipériodiques des cordes vocales forment l'essentiel de l'excitation du conduit vocal en parole voisée. Parmi les événements qui caractérisent le cycle du larynx, les instants d'ouverture et de fermeture de la glotte occupent une place importante. Ces instants sont notés GCI pour la fermeture (Glottal Closure Instant) et GOI pour l'ouverture (Glottal Opening Instant). L'excitation majeure du conduit vocal se produit à l'instant de fermeture de la glotte. Les cordes vocales se ferment brusquement et restent fermées pour à peu près la moitié du cycle (phase fermée), l'écoulement d'air s'arrête rapidement, ceci entraîne une discontinuité de l'onde du débit glottique appelée aussi flux glottal. Le GCI représente l'instant le plus significatif lorsqu'un point de référence périodique est recherché, car c'est l'événement le plus simple à détecter dans une période du pitch. Tout aussi important est le GOI qui sépare la phase ouverte de la phase fermée mais de façon plus lente et régulière, c'est l'événement le plus difficile à repérer. La connaissance des instants d'ouverture et de fermeture permet de déterminer divers paramètres relatifs à la qualité du son dont principalement la période du pitch T_0 et le quotient ouvert O_q durée de la phase ouverte relative à la période du signal. Le traitement de parole utilise les GCIs et les GOIs ; par exemple, lors de la phase fermée de la glotte, pour déterminer avec précision les paramètres du conduit vocal, tels que les formants [1]. De même l'analyse pitch synchrone nécessite naturellement une connaissance robuste des instants d'ouverture et de fermeture de la glotte. Dans les applications de synthèse vocale, les performances des méthodes PSOLA [2] dépendent fortement de la précision de localisation des marqueurs de pitch. Enfin il est également possible de disposer d'une meilleure caractérisation de la source par une analyse du signal qui se réfère aux GCIs et GOIs. Le signal d'excitation du conduit vocal est porteur d'informations caractéristiques sur la source vocale, l'analyse de ce signal permet de repérer aisément les instants d'ouverture et de fermeture de la glotte. Cette simplicité devient une tâche ardue si l'on cherche à détecter ces instants sur le signal de parole, en effet l'interaction de l'excitation glottique et les variations temporelles de la forme du conduit vocal rendent le signal complexe. Il est ainsi difficile d'isoler l'excitation majeure dans une période du pitch sur le signal de parole. Généralement, dans une période du pitch, on peut noter la présence de plusieurs excitations dont certaines sont significatives. Même si l'instant de fermeture de la glotte correspond à l'excitation majeure dans le cas d'une voyelle normale, en faible voisement, cet instant devient difficile à définir. A l'instant d'ouverture, l'excitation est de nature plus douce donc plus difficile à repérer et pratiquement impossible à détecter sur le signal de parole [1]. Nombreuses, sont les recherches qui tentent d'identifier avec la meilleure précision possible, les instants de fermeture de la glotte GCIs sur le signal d'excitation et le signal de parole, afin de disposer d'une référence périodique et d'une estimation de la période instantanée du pitch. Toutefois, les travaux sur l'instant d'ouverture sont relativement modestes. Ceci s'explique d'une part par la position d'importance première accordée à l'instant de fermeture et d'autre part aux difficultés associées à la détermination de l'instant d'ouverture. Les méthodes proposées pour la détection des instants de fermeture de la glotte à partir du signal de parole sont essentiellement basées soit sur une modélisation LPC de l'erreur comme le résidu ou le filtrage de Kalman [3]; soit sur les discontinuités spectrales observées sur les représentations temps fréquence ou temps échelle. La détection des instants d'ouverture est plus aisément obtenue sur le signal électroglottogramme (EGG) [4], cette approche n'est pas commode dans les applications [5].

Durant ces dernières décennies, la transformée en ondelettes a été fréquemment utilisée pour la détection des singularités d'un signal [6]. La transformée en ondelettes a été aussi utilisée par différents algorithmes pour la détection du pitch [7], [8]. De même l'analyse des coefficients en ondelettes a permis la détermination des instants de fermeture de la glotte [9]. En effet, les instants de fermeture de la glotte correspondent la plupart du temps, à des variations brusques ou des singularités dans le signal de parole.

Ce papier présente une approche simple et robuste de détection des instants d'ouverture et de fermeture de la glotte basée sur le produit des coefficients de la transformée en ondelettes continue du signal à différentes échelles. Nous avons montré dans des travaux antérieurs, que la transformée en ondelettes donne des résultats intéressants sauf dans certains cas où les singularités sont indiscernables sur les coefficients de la transformée en ondelettes [9]. En effet, le signal de parole présente des singularités lissées aux instants d'ouverture et de fermeture de la glotte [10]. Il est ainsi assez difficile dans certains cas de détecter ces instants sur le signal de parole. Pour circonvenir à ce problème, il est alors judicieux d'opérer une combinaison non linéaire des coefficients de la transformée en ondelettes à différentes échelles afin de donner une meilleure estimation des instants d'excitation de la source.

2. MODULES MAXIMA DE LA TRANSFORMEE EN ONDELETTES

Ce paragraphe montre que les coefficients de la transformée en ondelettes présentent des modules maxima aux singularités du signal. Nous allons d'abord montrer que la transformée en ondelette agit comme un opérateur différentiel multiéchelle d'ordre n, lorsque l'ondelette possède n moments nuls [6]. La transformée en ondelettes est alors la dérivée $n^{ième}$ de la fonction f lissée par une fonction.

La transformée en ondelettes d'une fonction f(t) peut se mettre sous la forme d'un produit de convolution

$wf(u, s) = f * \overline{\psi}_s(u)$, ψ étant l'ondelette mère, avec $\overline{\psi}_s(t)=\frac{1}{\sqrt{s}}\psi\left(\frac{-t}{s}\right)$. Dans laquelle une ondelette à n moments nuls s'écrit comme la dérivée nième d'une fonction de lissage θ ; La transformée en ondelettes s'écrit alors:

$wf(u, s) = s^n \frac{d^n}{du^n}(f * \overline{\theta}_s)(u)$, Le signal est lissé par la fonction $\overline{\theta}_s$.

Comme la transformée en ondelettes de f peut s'écrire comme un opérateur différentiel multiéchelle de f convolué par une fonction de lissage appropriée à l'échelle, lorsque l'ondelette n'a qu'un seul moment nul, les modules maxima de $|Wf(u,s)|$ correspondent aux maximums de la dérivée première de la fonction lissée. Quand la transformée en ondelettes de f n'a pas de maximum local aux fines échelles, alors f est localement régulière. Les singularités se trouvent ainsi au niveau des abscisses où convergent les modules maxima d'ondelettes aux fines échelles, et sont caractérisées par l'ordre de l'ondelette.

3. METHODE DU PRODUIT MULTIECHELLE (MPM)

Le produit des coefficients de la transformée en ondelettes à travers un nombre d'échelle a été fréquemment utilisé en traitement d'image. Pour caractériser les maxima associés aux véritables discontinuities du signal, Mallat et Zhong [11] analysent les propriétes des modules maxima des singularités à travers les échelles. La dérivée de Gaussienne et la spline quadratique sont les ondelettes particulièrement utilisées dans ce cas. La simple multiplication des transformées en ondelettes à des échelles adjacentes, fait ressortir les discontinuités [12], [13], [14]. Sadler et Swami [12] ont étudié l'efficacité de la méthode du produit multiéchelle en présence de bruit. Nous proposons d'utiliser la méthode du produit multiéchelle (MPM). Cet algorithme procède au calcul du produit des coefficients de la transformée en ondelettes pour différentes échelles dyadiques successives. Dans cette opération non linéaire sur f, les singularités produisent à travers les échelles des pics observés sur les coefficients de la transformée en ondelettes qui sont renforcés par le produit multiéchelle. La combinaison non linéaire tend à rehausser les maxima en atténuant les faux pics. Les pics de la transformée vont s'aligner à travers les échelles et non pas pour toutes les échelles parce qu'en augmentant l'effet du lissage, les singularités voisines vont interférer. Ainsi le choix d'une valeur d'échelle assez élevée, fait perdre l'alignement des pics dans le produit multiéchelle. Le nombre impair des termes du produit permet de préserver le signe de la singularité. Trois échelles dyadiques consécutives suffisent généralement pour la détection des pics par le produit.

4. APPLICATION DE MPM SUR LE SIGNAL DE PAROLE

L'ondelette utilisée dans ce travail est la dérivée première de gaussienne. Cette ondelette assure la détection des discontinuités présentes sur le signal de parole; de plus cette ondelette assure la propagation des maxima aux fines échelles [6].

La base de sons de l'Université de Keele est utilisée. Elle est composée de 5 voix féminines et 5 voix masculines. La base contient le signal de parole et le signal EGG. Les deux signaux sont échantillonnés à la cadence 20 kHz. Le signal EGG permet de donner les

instants références des GCIs et GOIs. En effet la dérivée du signal EGG montre deux sortes de pics opposés; les grands pics et qui sont des minimums, concordent à la fermeture de la glotte et les petits pics qui sont des maximums, correspondent à l'ouverture glottique [4].

La figure 1 montre le signal de parole, les coefficients de la transformée en ondelettes à trois échelles, le produit multiéchelle (PM) et enfin la dérivée du signal EGG prise comme référence. Les lignes de maxima de plus forte amplitude correspondent aux instants de fermeture de la glotte repérés sur le signal DEGG. Les GOIs ne sont pas discernables ici sur les transformées en ondelettes.

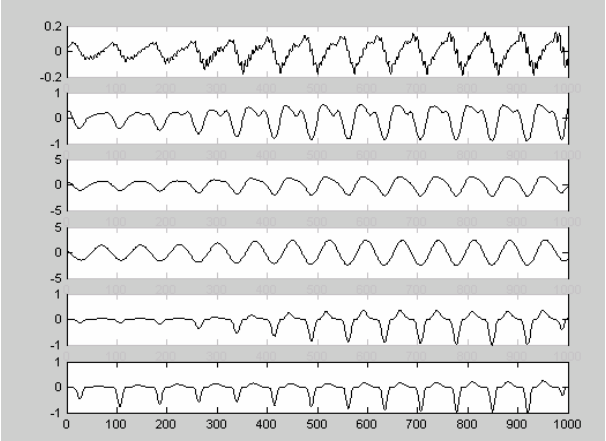

Figure 1 : Signal de parole, locuteur f4, voyelle /o/, mot /north/,3 échelles de transformées en ondelettes, le PM et le DEGG.

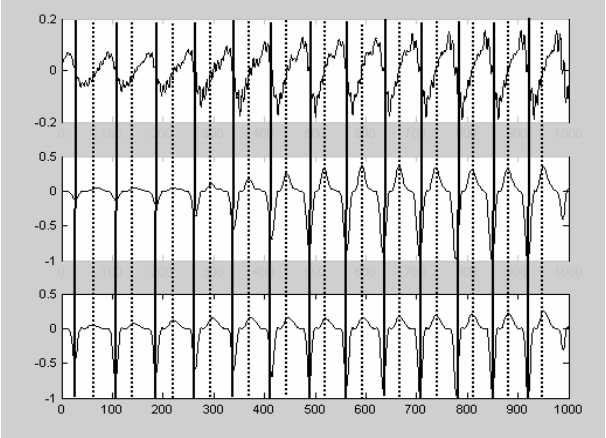

Figure 2 : Signal de parole locuteur f4, PM, DEGG.

La figure 2 montre une tranche d'une zone voisée du locuteur f4 suivie du produit multiéchelle du signal de parole et du signal DEGG. Le produit montre alors deux types d'extrema ; des minimums de forte amplitude marqués par des traits continus et qui correspondent aux instants de fermeture de la glotte et des maximums de moins forte amplitude marqués par des traits interrompus et qui coïncident avec les instants d'ouverture de la glotte donnés par le signal DEGG. Dans ce cas la coïncidence est parfaite.

La figure 3 donne un second exemple : signal de parole, PM et DEGG pour le locuteur f1. Notons également la parfaite correspondance entre les extrema fournis par le produit multiéchelle et ceux donnés par le signal DEGG aussi bien pour l'ouverture que la fermeture de la glotte.

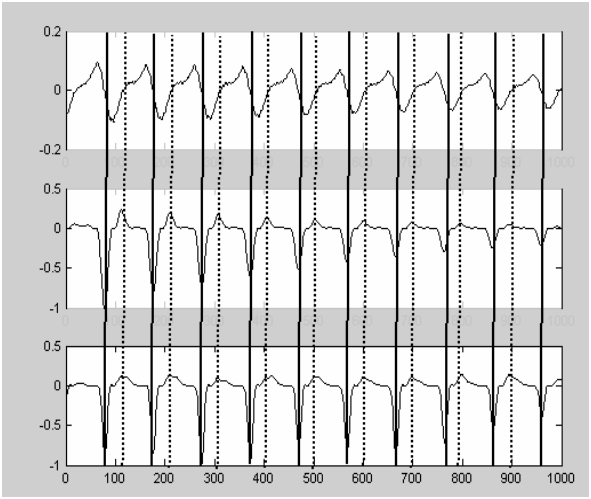

Figure 3 : Signal de parole locuteur f1, PM, DEGG.

Le calcul du produit des coefficients de la transformée en ondelettes du signal de parole avec l'ondelette dérivée première de gaussienne aux trois premières échelles dyadiques, permet une détection précise des GCIs et des GOIs grâce aux pics qui deviennent plus marqués à la fermeture de la glotte. Et l'apparition de pics à l'ouverture. Le signal issu du produit présente une allure semblable au signal DEGG.

5. EVALUATION DE LA METHODE MPM

Cette approche a été évaluée sur la base de sons de l'université de Keele. Nous avons procédé à la comparaison des valeurs instantanées des GCIs et GOIs références donnés par le signal DEGG et des GCIs et GOIs mesurés par la méthode du PM. Les valeurs obtenues dont les erreurs sont inférieures à 0.25 ms par rapport aux valeurs de références sont considérées correctes. L'erreur comprise entre 0.25 ms et 1 ms est comptée comme erreur fine. Les erreurs dépassant 1 ms sont comptées comme grosses. Sur la totalité des mesures des GCIs et GOIs références, nous déterminons le pourcentage des mesures estimées correctes (MC), le pourcentage de mesures correspondant aux erreurs fines (EF), le pourcentage de mesures correspondant aux erreurs grosses (EG), le pourcentage des mesures manquées (MM : cas où le PM ne montre pas un pic au GCI) et le pourcentage de fausses alarmes (FA : cas où le produit montre plus qu'un pic pour un GCI référence donné). L'évaluation est appliquée sur les segments voisés donnés comme tels par la base de sons. Les tableaux 1 et 2 présentent respectivement les performances de la méthode du produit pour la détection des GCIs et des GOIs pour les locuteurs hommes et femmes de la base.

Table 1 : Résultats d'évaluation de la méthode du produit pour la détection des GCIs.

	MC	EF	EG	MM	FA
Hommes	93.73	2.79	1.95	1.51	7.89
Femmes	96.67	1.41	0.22	1.67	0.51
La base	95.70	1.87	0.79	1.62	2.95

Table 2 : Résultats d'évaluation de la méthode du produit pour la détection des GOIs.

	MC	EF	EG	MM	FA
Hommes	49.03	16.36	11.01	23.59	3.38
Femmes	86.12	7.96	1.17	4.76	0.40
La base	73.67	10.78	4.47	11.08	1.4

La méthode proposée est plus performante pour la détection des instants de fermeture de la glotte que pour les instants d'ouverture. Toutefois, elle donne des résultats satisfaisants au regard de l'existant. Les deux tableaux montrent que la méthode proposée est globalement plus performante pour les femmes que pour les hommes aussi bien pour la détection des GCIs que des GOIs.

6. CONCLUSION

La détection des instants de fermeture et d'ouverture de la glotte est opérée à travers l'estimation des discontinuités du signal de parole. Les points caractéristiques du signal de parole sont déterminés par la méthode du produit multiéchelle. Cette méthode consiste à calculer les coefficients de la transformée en ondelettes, le produit de ces coefficients puis détecter les minima et maxima pour localiser respectivement les GCIs et GOIs. Le produit de trois échelles garantit la conservation du signe de la discontinuité. Le produit permet de renforcer les pics aux instants de fermeture de la glotte le long des échelles et de réduire les faux pics particulièrement aux instants d'ouverture glottique. L'évaluation de cette approche sur la base de Keele a donné un taux de mesures correctes de l'ordre de 96 % pour la détection des GCIs et un taux de 74% pour le repérage des GOIs. Ces résultats comptent parmi les meilleurs obtenus à ce jour.

BIBLIOGRAPHIE

[1] B. Yegnanarayana et R. N. J. Veldhuis. Extraction of Vocal-Tract System Characteristics from Speech Signals. *IEEE Trans. Speech and Audio Signals Proc.*, volume 6, pages 313-337, 1998.

[2] J.G. McKenna. Automatic glottal closed-phase location and analysis by Kalman filtering. in Proc. ISCA 2001, Tutorial and Research Workshop on Speech Synthesis, 2001.

[3] C. Hamon, E. Moulines and F. Charpentier. A Diphone Synthesis System Based on Time Domain Prosodic Modifications of Speech. In Proc. *Intl. Conf. Acoustics, Speech and Signal Processing*, pages 238-241, 1989.

[4] N. Henrich, C. d'Alessandro, M. Castellengo et B. Doval. Mesures électroglottographiques du quotient d'ouverture glottique en voix parlée et chantée. In *Proc. 23èmes Journées d'Etude sur la Parole*, 2000.

[5] N. Henrich. *Etude de la Source Glottique en Voix Parlée et Chantée: Modélisation et Estimation, Mesures Acoustiques et Electroglottographiques*. Phd Thesis, Paris 6 University, 2001.

[6] S. Mallat. *Une Exploration des Signaux en Ondelettes*. Les Editions de l'Ecole Polytechnique, Paris, 2000.

[7] L. Janer, J.J Bonet et E.L Lleida-Solano. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. *In Proc. Intl. Conf. on Spoken Language Processing*, 1996.

[8] S. Kadambe et G.F. Bourdeaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. On Information Theory*, volume 38, 1992.

[9] A.B. Slimane, A. Bouzid et N. Ellouze. Wavelet Decomposition of Voiced Speech and Mathematical Morphology Analysis for Glottal Closure Instants Detection. *In Proc. European Signal Processing Conference*, volume 3, pages 81-84, 2002.

[10] A. Bouzid et N. Ellouze. Caractérisation des Singularités du Signal de Parole. Colloque sur le traitement du signal et des images GRETSI, Tome 1, pages 273-276, 2003.

[11] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 14, pages 710-732, 1992.

[12] B.M. Sadler and A. Swami. Analysis of multiscale products for step detection and estimation. *IEEE Trans. Information Theory*, vololume 45, pages 1043-1051, 1999.

[13] B.M. Sadler, T. Pham and L.C. Sadler. Optimal and wavelet-based shock wave detection and estimation. *Journal of Acoustical Society of America*, volume 104, pages 955-963, 1998.

[14] P. Bao, L. Zhang and X. Wu. Canny. Edge detection enhancement by scale multiplication. *IEEE Trans. on Pattern Analysis and Machine Intelligenc*e, vol. 27, no. 9, pages 1485-1490, 2005.