Algorithme génétique pour le rehaussement d'un signal de parole voisé fondé sur un modèle sinusoïdal

E. GRIVEL[1], J. M. VESIN[2] et M. NAJIM[1]

[1]Equipe Signal et Image, UMR 5131 LAP/CNRS, ENSEIRB
B.P. 99, F- 33402 Talence Cedex, France,
e-mail : eric.grivel@tsi.u-bordeaux.fr, najim@tsi.u-bordeaux.fr

[2]Signal Processing Laboratory, Ecole Polytechnique Fédérale de Lausanne
CH 1015 Lausanne, Suisse,
e-mail : Jean-Marc.Vesin@epfl.ch

Résumé : Parallèlement aux approches par atténuation spectrale à court terme, de nombreuses méthodes paramétriques ont été mises en œuvre pour rehausser le signal de parole, capté par un unique microphone. Ces méthodes génèrent habituellement un bruit résiduel, de nature musicale, plus ou moins perceptible à l'écoute du signal rehaussé. Ici, nous proposons de restaurer un signal de parole voisé, en utilisant un modèle de type sinusoïdal. Deux modèles sont traités : celui d'une somme de fonctions de Gabor réelles et celui d'une somme de sinusoïdes. La procédure d'estimation des paramètres repose sur un algorithme de type « matching pursuit », utilisant les algorithmes génétiques. Un détecteur d'activité vocale n'est en outre pas nécessaire pour estimer la variance du bruit additif. Le signal rehaussé obtenu est alors comparable à celui obtenu par les méthodes d'atténuation spectrale à court terme.

Mots clés : rehaussement, parole, modèle sinusoïdal, algorithme génétique, matching pursuit.

Astract : In addition to traditional non-parametric methods using short time spectral attenuation, several parametric methods have been developed to enhance a speech signal, when a single sequence of noisy signal is available. Usually, when listening to the enhanced signal, the so-called residual musical noise is perceptible. Here, we propose to retrieve a voiced speech by using a sinusoidal model. Two models are considered: the first one is based on a sum of real Gabor functions whereas the second is a sum of cosines. Estimating the model parameters is completed by using a matching-pursuit based method requiring Genetic Algorithms (GA). It should be noted that our approach does not require a Vocal Activity Detector (VAD) to estimate the variance of the additive noise, which is a great advantage. In addition, the weak residual noise is comparable to the one obtained with traditional non-parametric methods using short time spectral attenuation.

Keywords : enhancement, speech, sinusoidal model, genetic algorithm, matching pursuit.

1. Introduction

Les approches paramétriques sont de nos jours très largement employées pour l'analyse et le traitement de la parole. Différents modèles décrivant l'évolution temporelle du signal de parole ont ainsi été étudiés au cours de ces quarante dernières années.

Le modèle source-filtre de type AR s'est largement répandu en codage (codeur CELP et ses dérivés) du fait du nombre réduit de paramètres ainsi qu'en synthèse de la parole. Dans le contexte du rehaussement de la parole, les méthodes fondées sur un modèle autorégressif excité par un bruit blanc gaussien ont été aussi largement étudiées dans les publications ([5], [6], [8], [17] par exemple). Ces approches nécessitent en général l'utilisation d'un filtre de Kalman, voire d'un lissage de Kalman pour réduire le bruit résiduel musical. Des résultats significatifs ont été obtenus bien que le modèle autorégressif excité par un bruit blanc ne soit pas bien adapté à des trames dites voisées, c'est-à-dire présentant une nature quasi périodique. Pour cette raison, Goh et al. [7] ont proposé d'adapter l'excitation (soit périodique, soit blanche) à la nature de la trame analysée, à condition que la décision de voisement/non voisement ait été préalablement prise et la valeur du pitch estimée.

Pour un signal voisé ou sonore, la représentation sinusoïdale (harmonique) a été souvent utilisée pour la synthèse de la parole. Elle a pour objet de décomposer le signal en une somme de segments sinusoïdaux (resp. harmoniques) [18]. On peut noter que, pour formaliser à la fois les natures voisée et non voisée de la parole, la superposition d'une composante harmonique et d'une composante de bruit a été traitée [21], [16] et a abouti à une modélisation dite « harmonique + bruit », essentiellement testée en synthèse. Cependant, dans le contexte du débruitage d'un signal de parole, peu de travaux fondés sur une modélisation sinusoïdale ont été menés. Jensen et Hansen [11] ont proposé une remise à jour, trame par trame, de l'estimation des amplitudes réelles de chaque piste fréquentielle, à partir d'un filtre de Wiener ; le suivi des composantes fréquentielles s'effectue à partir d'un lissage de l'enveloppe spectrale de chaque trame analysée. Dans [9], les auteurs ont étudié un modèle sinusoïdal stochastique dont l'amplitude réelle de chaque « piste » est modélisée par un processus autorégressif d'ordre 1. Néanmoins, la plupart des travaux se fonde sur la modélisation du signal de parole par une somme d'exponentielles complexes ; ainsi, Jensen et al. [10] proposent d'exploiter les techniques de décomposition en sous-espaces signal et bruit pour débruiter un signal de parole perturbé par un bruit blanc. Cette approche permet alors d'extraire les premiers formants du signal à partir des observations bruitées, mais le signal rehaussé présente un bruit résiduel de nature musicale. Récemment, une approche réduit ce phénomène indésirable en lissant le spectre du signal rehaussé à partir d'une estimation du masque psycho acoustique fréquentiel [12].

Dans cette communication, nous proposons d'utiliser un modèle de type sinusoïdal réel pour le rehaussement d'un signal de parole voisée. Deux types de fonctions sont envisagés : soit des fonctions sinusoïdales, soit des fonctions de Gabor réelles. La base de fonctions caractérisant le signal est obtenue en mettant en œuvre une approche de type « matching pursuit » [14]. L'originalité de ce travail vient de l'utilisation des algorithmes génétiques qui permettent d'estimer les paramètres caractérisant chaque fonction. Ce type d'approche a été déjà utilisé dans une large gamme d'applications, notamment dans le contexte biomédical pour la détection de formes particulières dans des électroencéphalogrammes ou des électrocardiogrammes [22], en imagerie médicale [19], en rehaussement d'images, etc.

Cette communication s'organise comme suit : dans la partie 2, nous présentons les deux modèles du signal de parole voisée que nous avons adoptés. Dans la partie 3, nous introduisons les algorithmes génétiques et présentons les différentes étapes de l'approche de rehaussement du signal de parole. Enfin, nous présentons une étude comparative fondée sur le gain du rapport signal à bruit (RSB) pour différentes valeurs du RSB d'entrée (5 à 15dB) et l'écoute des signaux rehaussés. Nous verrons en outre si cette approche paramétrique peut supplanter les méthodes d'atténuation spectrale à court terme [2], [3], [4] [13], [1], [15].

2. Deux modèles pour le signal de parole voisé

Le rehaussement du signal de parole voisé $s(k)$ perturbé par un bruit additif $b(k)$ est effectué sur des trames de N échantillons (à la fréquence d'échantillonnage f_e, soit une durée de la trame de l'ordre de 10ms) avec un fenêtrage de Hamming et un recouvrement de 50%. Nous proposons alors d'utiliser un des deux modèles de type sinusoïdal suivants :

Le premier consiste en une somme de L « pistes » sinusoïdales. L peut varier d'une trame à l'autre. On a alors :

$$y(k) = s(k) + b(k) \tag{1}$$

$$s(k) = \sum_{l=1}^{L} a_l \cos(\varphi_l(k)). \tag{2}$$

La phase de chaque piste sinusoïdale est de la forme :

$$\varphi_l(k) = 2\pi f_l \left(\frac{k - k_{0,l}}{f_e} \right) + \beta_l \tag{3}$$

Puis, le signal est modélisé par une somme de fonctions de Gabor réelles. On obtient alors :

$$s(k) = \sum_{l=1}^{L} c_l \exp\left(-\frac{1}{2} \left(\frac{k - k_{0,l}}{f_e {\sigma_l}^2} \right) \right) \sin(\varphi_l(k)). \tag{4}$$

3. Matching pursuit fondé sur les algorithmes génétiques (AG)

3.1 Introduction aux AG

Les algorithmes génétiques sont inspirés du concept de sélection naturelle. Bien adaptés aux problèmes d'optimisation, ils consistent à générer une population aléatoire « d'individus » (c'est-à-dire un ensemble de solutions potentielles) et à la faire évoluer de manière itérative jusqu'à ce qu'un critère dit d'arrêt soit vérifié. Les AG conservent dans la population testée les individus qui maximisent une fonction d'évaluation (« fitness », en anglais) pour constituer une génération intermédiaire (« mating pool ») ; c'est l'étape dite de reproduction. Les autres individus sont remplacés en utilisant des règles d'évolution fondées sur différents opérateurs génétiques (croisement, mutation, etc.). Cf. figure 1. Pour plus de détails, nous invitons le lecteur à se référer à [23].

3.2 AG pour le débruitage de la parole

La procédure de rehaussement de la parole repose ici sur un algorithme itératif de type « matching pursuit » fondé sur l'utilisation des AG.

Un « individu » correspond à un vecteur θ concaténant les paramètres d'une piste fréquentielle du modèle du signal de parole. Ainsi, pour le modèle proposé en (2) -resp. (4)-, on considère le vecteur $\theta = [k_{0,l}, \sigma_l, f_l, \beta_l]$ -resp. $\theta = [k_{0,l}, f_l, \beta_l]$-. A noter que l'on contraint chaque paramètre dans un intervalle donné. L'angle β_l est compris entre 0 et 2π.

A l'aide des algorithmes génétiques, on sélectionne la piste fréquentielle qui approxime au mieux la trame analysée, au sens des moindres carrés. Cela revient à déterminer le vecteur θ qui minimise le carré du module du résidu de la projection de la trame sur la "piste" fréquentielle.

Selon la nature du résidu (blanche ou non), l'extraction d'une nouvelle piste fréquentielle n'est pas nécessaire. Si le résidu est blanc, le signal rehaussé correspond à la somme des différentes pistes fréquentielles estimées ; sinon, la procédure d'extraction de pistes se poursuit et consiste alors à approximer, au sens des moindres carrés, le résidu.

Le test de blancheur [20] que nous choisissons pour le résidu vérifie :

$$\left| \gamma_{rr}(k) \right| \leq 1{,}95 \frac{\left| \gamma_{rr}(0) \right|}{\sqrt{N}} \text{ pour } k > 0 ,$$

où $\gamma_{rr}(k)$ dénote la fonction d'autocorrélation du résidu de la projection.

A noter qu'une telle approche de rehaussement ne nécessite pas l'estimation de la variance du bruit additif, contrairement aux méthodes d'atténuation spectrale à court terme [2], [3] ou de certaines méthodes paramétriques [7], [6], [10].

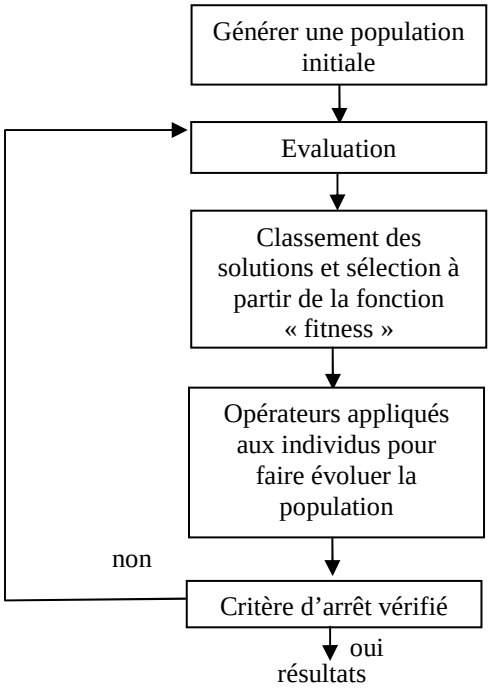

FIG 1 : organigramme d'un algorithme génétique.

3.3. Améliorations de la méthode et commentaires

Améliorations n°1 :

Pour réduire la complexité calculatoire et éviter le test de blancheur, on peut considérer un nombre limité prédéterminé de pistes *L*.

Améliorations n°2 :

Si l'on implante l'approche présentée en 3.2, la méthode fournit des signaux rehaussés mais entachés d'un bruit résiduel. Ce phénomène est essentiellement dû au choix de l'ordre du modèle et en particulier à sa surestimation pour différentes trames. En effet, certaines pistes fréquentielles, de faible énergie, peuvent être extraites à l'aide des algorithmes génétiques. Elles sont alors uniquement prises en compte pour la trame courante mais n'apparaissent pas nécessairement dans la précédente ni la suivante, ce qui génère un bruit de nature musicale.

Pour pallier ce phénomène indésirable, nous proposons, une fois les *L* pistes fréquentielles extraites, d'introduire un critère énergétique fondé sur la variance du bruit additif σ_b^2. Si la puissance d'une piste fréquentielle vérifie : $P_{piste\ fréquentielle} \leq \alpha\sigma_b^2$, où α est un hyper paramètre, la composante n'est pas prise en compte (Critère noté Amé. 2a/). Cependant cette démarche nécessite un détecteur d'activité vocale (DAV) pour estimer pendant les trames de silence la variance du bruit additif, supposé stationnaire.

Afin d'éviter l'utilisation d'un DAV, nous proposons de considérer la variance du résidu final au lieu de σ_b^2 (Critère noté Amé. 2b/).

4. Résultats

Le signal voisé est échantillonné à 16 kHz. Les méthodes sont testées pour des rapports signaux sur bruits (RSB) en entrée compris entre 5 et 15 dB.

RSB entrée	Modèle (2) + *Amé. 2a*	Modèle (4) + *Amé. 2a*	Modèle (2) + *Amé. 2b*	Modèle (4) + *Amé. 2b*
5 dB	6,2 dB	4,5 dB	6,2 dB	5,3 dB
10 dB	5,1 dB	3,8 dB	5,5 dB	4,5 dB
15 dB	3,2 dB	2,3 dB	3,6 dB	2,8 dB

TAB 1 : gains moyens du RSB pour différentes valeurs du RSB en entrée.

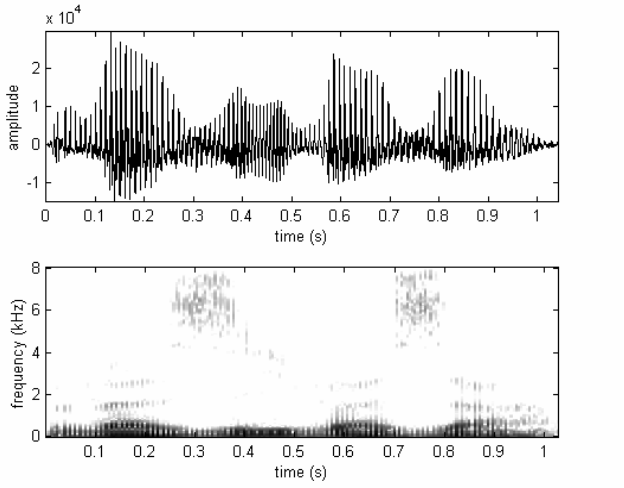

FIG 2 : représentation temporelle et spectrogramme du signal original

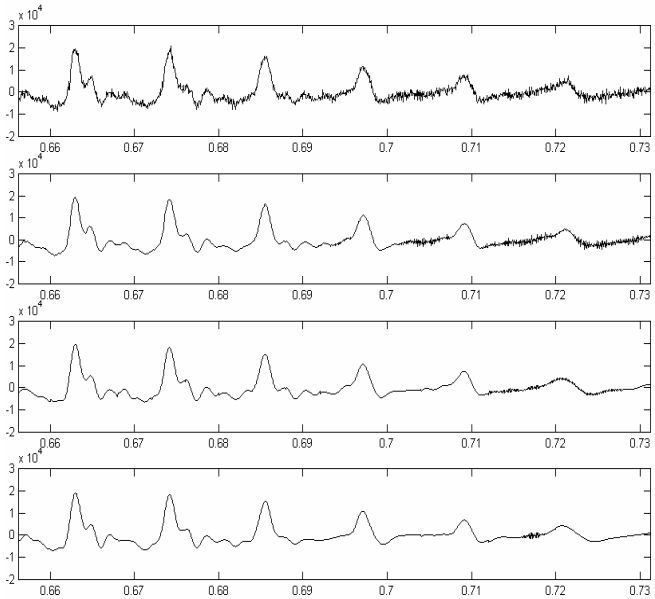

FIG 3 : zoom sur la représentation temporelle des signaux bruité (relevé 1), original (relevé 2) et des signaux rehaussés respectivement avec le modèle (2) (relevé 3) et le modèle (4) (relevé 4), avec un RSB en entrée de 15 dB.

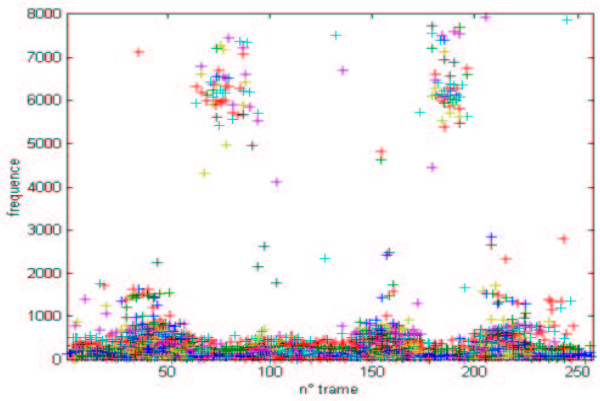

FIG 4 : représentation au cours des trames des fréquences f_l estimées pour le modèle (2) pour un RSB d'entrée de 15 dB.

Les différences de gain en RSB qui existent entre les différentes méthodes proposées sont essentiellement dues au choix de α. Il répond à un compromis entre la limitation du bruit musical et une distorsion du signal. De plus, des différences pour l'estimation de fricatives telles que les /z/ sont à prendre en compte (Cf. Fig. 3).

Le signal rehaussé à partir du modèle (4) est plus « lissé » que celui obtenu à partir de (2). En outre, à l'écoute du signal rehaussé (pour un RSB d'entrée de 10 ou 15 dB par exemple), le bruit résiduel n'est pas perceptible et comparable à celui obtenu avec [4].

5. Conclusion

L'approche paramétrique proposée ici permet d'obtenir un signal rehaussé comparable à celui obtenu avec [4]. Dans sa version améliorée, elle a l'avantage de ne pas nécessiter de DAV. Dans le cas d'une extension au cas coloré, l'introduction d'une étape de pré-blanchiment pourrait être envisagée. Enfin, un traitement à complexité réduite pourrait être mis en œuvre.

Références

[1] A. Akbari Azirani, Rehaussement de la parole en ambiance bruitée, application aux télécommunications mains libres. Univ. de Rennes I. Nov. 1995.

[2] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of Speech Corrupted by Acoustic Noise, ICASSP, pp. 208-211, April 1979.

[3] S. F. Boll, Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Trans. Acoust., Speech, Signal Processing, vol. 27, n°2, pp. 113-120, 1979.

[4] Y. Ephraim and D. Malah, Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, IEEE Trans. Acoust., Speech, Signal Processing, vol. 32, n°6, pp. 1109-1121, 1984.

[5] M. Gabrea, E. Grivel and M. Najim, A Single Microphone Kalman Filter-Based Noise Canceller, IEEE Signal Processing Letters, pp. 53-55, march 1999.

[6] J. D. Gibson, B. Koo and S. D. Gray, Filtering of Colored Noise for Speech Enhancement and Coding, IEEE Trans. on Signal Processing, vol. 39, n°8, pp. 1732-1742, August 1991.

[7] Z. Goh, K.-C. Tan and B. T. G. Tan, Kalman-Filtering Speech Enhancement Method Based on a Voiced-Unvoiced Speech Model, IEEE Trans. On Speech and Audio Processing, vol. 7, no. 5, pp. 510-524, 1999.

[8] E. Grivel, M. Gabrea, et M. Najim, Speech Enhancement as a realization issue, Signal Processing, vol 82, n°12, pp 963-1978, Dec. 2002.

[9] E. Grivel, A. Ferrari and O. Cappé, Le modèle harmonique stochastique et son application au rehaussement de signal de parole, GRETSI '01, 10-14 Sept 2001.

[10] S. H. Jensen, P. C. Hansen, S. D. Hansen and J. Sorensen, Reduction of Broad Band Noise in Speech by Truncated QSVD, IEEE Trans. On Speech and Audio Processing, vol. n°3, n°6, 1995, pp. 439-448.

[11] J. Jensen and J. L. Hansen, Speech Enhancement Using a Constrained Iterative Sinusoidal Model, IEEE Trans. On Speech and audio Processing, vol. 9, n°7, Oct. 2001.

[12] M. Klein and P. Kabal, Signal Subspace Speech Enhancement with Perceptual Post Filtering, ICASSP, vol. 1, pp 537-540, 2002.

[13] R. Le Bouquin and G. Faucon, "Maximum Likelihood Noise Cancellation With Spectral Constraints", ICASSP 91, Toronto, Canada, pp. 941-944, April 1991.

[14] S. G. Mallat and Z. Zhang, "Matching Pursuits With Time-Frequency Dictionaries", IEEE Trans. On Signal Processing, vol. 41, n°12, Dec. 1993.

[15] R. J. Mc Aulay and M. L. Malpass, Speech Enhancement Using a Soft-Decision Noise Suppression Filter, IEEE Trans. on Audio, Signal and Speech Processing, vol. 28, n°2, pp. 137-145, April 1980.

[16] M. Oudot, Application du modèle Sinusoïudes et bruit au décodage, au débruitage et à la modification des sons de parole, Thèse, ENST, Paris, 1998.

[17] K. K. Paliwal and A. Basu, A Speech Enhancement Method Based on Kalman Filtering, ICASSP 87, pp. 177-180.

[18] T.F. Quatieri and R.J McAulay, Noise reduction using a soft-decision sine-wave vector quantizer, ICASSP, Albuquerque, New Mexico, 1990.

[19] P. Shroeter, J. M. Vesin, T. Langenberger and R. Meuli, Robust Parameter Estimation of Intensity Distributions for Brain Magnetic Resonance Images, IEEE Trans on Medical Imaging, Vol. 17, n°2, pp172-186, April 1998.

[20] P. Stoica, A Test for Whiteness, IEEE Trans. on Automatic Control, vol. AC-22, pp. 992-993, 1977.

[21] Y. Stylianou, Modèles harmoniques plus bruit combinés avec les méthodes statistiques, pour la modification de la parole du locuteur, Thèse, ENST, 1996.

[22] K. Suárez, J. Silva, M. Najim. A Genetic Algorithm Approach For Pattern Recognition In Biomedical Signals, GRETSI 2001, 10-14 Sept 2001.

[23] J. M. Vesin, Efficient Implementation Of Matching Pursuit Using A Genetic Algorithm In The Continuous Space, EUSIPCO 2000.