Estimation de la source glottique par décomposition modale empirique

Kemiha Mina[1] Abdellah Kacha[1], Mounir Boudjerda[1,2]
(1) Laboratoire de Physique de Rayonnement et Applications, Université de Jijel, Algérie
(2) Welding and NDT Research Centre (CSC), Alger, Algérie
kemihamina@yahoo.fr, kacha_a@yahoo.com, boudjerda@yahoo.com

RESUME

Dans cet article, la décomposition modale empirique est proposée comme alternative pour estimer la source glottique à partir du signal de parole. En utilisant l'algorithme de décomposition modale empirique, le logarithme de l'amplitude du spectre du signal de parole est décomposé en composantes oscillatoires appelées fonctions de modes intrinsèques. Une procédure adaptative est ensuite utilisée pour sélectionner les fonctions de modes intrinsèques appropriés qui constituent l'amplitude dans le domaine log spectral de la source glottique. L'exploitation de l'information de phase obtenue à partir du signal acoustique conjointement avec la somme des fonctions de modes sélectionnées permet d'obtenir une estimation de la source glottique. La méthode proposée est testée sur des signaux de parole synthétiques et comparée avec la méthode d'estimation de la source glottique basée sur le cepstre.

ABSTRACT

In this paper, the empirical mode decomposition algorithm is proposed as an alternative to estimate the glottal source from the speech signal. Using the empirical mode decomposition algorithm, the logarithm of the magnitude spectrum of the speech signal is decomposed into oscillatory modes names intrinsic mode functions. An adaptive procedure is then used to select the appropriate intrinsic mode functions that constitute the magnitude of the glottal source in the log spectral domain. The exploitation of the phase information jointly with the sum of the selected intrinsic mode functions provides an estimate of the glottal source. The proposed method is tested on a synthetic speech signals and compared to the cepstrum-based glottal source estimation method.

MOTS-CLES : décomposition modale empirique, fonctions de modes intrinsèques, estimation de la source glottique.

1 Introduction

La séparation de la parole en ses deux contributions réponse fréquentielle du conduit vocal et excitation glottique est un sujet important en traitement de la parole. Isoler explicitement les deux composantes permet de les modéliser de façon indépendante. La caractérisation de l'excitation glottique présente plusieurs avantages en reconnaissance du locuteur (Plumpe et al. 1999), dans la caractérisation et l'analyse des troubles de la voix (Moore et al. 2003), en reconnaissance de la parole (Yamada, et al. 2002) et en synthèse de la parole (Drugman et al. 2009). Ces raisons justifient la nécessité de développer des algorithmes capables

d'estimer et de paramétrer le signal glottique de manière robuste et fiable.

Bien que les techniques de modélisation du conduit vocal soient assez bien établies, ce n'est pas le cas de la représentation de la source glottique. Certains travaux ont abordé le problème d'estimation de la contribution de la glotte directement à partir de la forme d'onde du signal de parole. La plupart des approches s'appuie sur une première modélisation paramétrique du conduit vocal, puis utilise le filtrage inverse afin d'éliminer l'effet du conduit vocal et obtenir une estimation du signal glottique. Dans (Alku et al. 1994), le modèle discret tout pôle a été utilisé pour modéliser le conduit vocal. La technique itérative adaptive par filtrage inverse décrit dans (Alku et al. 1992) isole le signal source en estimant de manière itérative à la fois les composantes dues au conduit vocal et à la source. Dans (Brookes et Chan, 1994), l'estimation du signal glottique est affinée sur plusieurs cycles glottiques. Une technique non paramétrique basée sur les zéros de la transformée en Z (ZZT) et le cepstre complexe a été proposée dans (Bozkurt et al. 2005 ; Doval et al. 2003). Cette approche repose sur l'observation que la parole est un signal à phase mixte comprenant une composante causale et une composante anti-causale où la composante anti-causale correspond à la phase d'ouverture de la glotte et la composante causale comprend à la fois la fermeture de la glotte et les contributions du conduit vocal.

Récemment, une méthode de décomposition du signal, appelée décomposition modale empirique (EMD), a été introduite pour analyser les données issues de processus non stationnaires et/ou non linéaires (Huang N.E. et al. 1998). La décomposition modale empirique a reçu plus d'intérêt en termes d'applications, d'interprétation et d'amélioration. L'avantage majeur de la décomposition modale empirique est que les fonctions de base sont obtenues à partir du signal lui-même et non fixées a priori comme dans les méthodes d'analyses conventionnelles (transformée de Fourier, transformée en ondelettes, etc.). Dans (Kacha et al. 2012 ; Kacha et al. 2013), la décomposition modale empirique a été proposée pour décomposer le logarithme du spectre du signal de parole en trois composantes qui sont la composante harmonique, la réponse fréquentielle du conduit vocal, et le bruit. Ces composantes ont été utilisées par la suite pour définir un indice acoustique appelé rapport harmonique sur bruit.

Dans cet article, la décomposition modale empirique est proposée comme alternative pour estimer la source glottique à partir du signal de parole. La méthode d'estimation proposée opère dans le domaine log-spectral. L'efficacité de l'approche proposée est évaluée sur la parole synthétique et sa performance est comparée à celle du cepstre complexe (Drugman et al. 2009 ; Drugman et al. 2012) dont l'utilisation impose un ensemble de contraintes sur la fenêtre d'analyse pour que le signal de parole puisse être modélisé par un modèle à phase mixte.

Le reste du papier est organisé comme suit. L'algorithme de décomposition modale empirique est présenté dans la section 2. L'approche d'estimation de la source glottique basée sur la décomposition modale empirique est présentée dans la section 3. Les résultats basés sur les signaux synthétiques sont présentés dans la section 4. Enfin, les conclusions sont données dans la section 5.

2 Décomposition modale empirique

La décomposition modale empirique est une méthode qui repose sur une décomposition adaptée en décrivant localement le signal comme une succession de contributions d'oscillations rapides (hautes fréquences) sur des oscillations plus lentes (basses fréquences) proposée par (Huang N.E. et al. 1998). En raison de sa nature empirique et algorithmique, il n'existe actuellement aucun fondement théorique complet de cette méthode. La méthode de décomposition modale empirique dispose d'un processus de tamisage fini. Les signaux peuvent être décomposés en une série de fonctions de modes intrinsèques (IMF) qui doivent satisfaire à deux conditions : (i) le nombre de passages par zéro et le nombre d'extrema sont égaux ou ne diffèrent pas plus d'un, (ii) la valeur moyenne de l'enveloppe produite par les extrema locaux est égale à zéro. Dans ce procédé, les fonctions de modes intrinsèques sont essentiellement obtenues à partir de points extrema locaux qui représentent des échelles de temps caractéristiques du signal. Au moyen de l'EMD, le signal à analyser est décomposé en N modes empiriques et un résidu. Les IMFs représentent les oscillations locales intrinsèques incrustées dans le signal, le mode résiduel peut être une fonction monotone ou d'énergie très faible peut être ignoré. Lorsque le mode résiduel satisfait à cette condition, les différentes étapes du procédé de tamisage peuvent s'écrire sous la forme du pseudo code suivant :

Etape 1 : initialisation $j \leftarrow 1$ $\left(j^{ième}\ IMF\right), r_{j-1}(t) \leftarrow x(t)$(résidu), Fixer le seuil Δ.

Etape 2 : extraire la $j^{ième}\ IMF$:

(a) $h_{j,i-1}(t) \leftarrow r_{j-1}(t),\ i \leftarrow 1$ (i, itération de la boucle de tamisage)

(b) extraire les maxima et minima locaux de $h_{j,i-1}(t)$

(c) calculer les enveloppes supérieures et inférieures : $U(t)$ et $L(t)$ par interpolation entre les maxima et les minima locaux $h_{j,i-1}(t)$ respectivement.

(d) calculer l'enveloppe moyenne : $\mu(t) \leftarrow (\ U(t) + L(t))/2$

(e) mettre à jour : $h_{j,i}(t) \leftarrow h_{j,i-1}(t) - \mu(t), i \leftarrow i + 1$.

(f) calculer le critère d'arrêt (par exemple) $SD = \sum_{t=0}^{T} \frac{\left|h_{j,i-1}(t) - h_{j,i}(t)\right|^2}{(h_{j,i-1}(t))^2}$

où T représente la durée totale du signal.

(g) décider : répéter l'étape (b)-(f) jusqu'à ce que $SD < \Delta$ et alors mettre $IMF_j(t) \leftarrow h_{j,i}(t)$ $(j^{ième} IMF)$.

Etape 3 : mettre à jour le résidu : $r_j(t) \leftarrow r_{j-1}(t) - IMF_j(t)$.

Etape 4 : répéter l'étape 2 avec $j \leftarrow j + 1$ jusqu'à ce que le nombre d'extrema dans $r_j(t)$ soit inférieur à 2.

Le processus de reconstruction du signal est donné par l'équation suivante,

$$x(t) = \sum_{j=1}^{N} IMF_j(t) + r_N(t), N \in N^*$$

3 Estimation de la source glottique

3.1 Principe

Selon le modèle source-filtre de production de la parole, le signal de parole peut être considéré comme le résultat de la convolution de l'excitation de l'appareil vocal (excitation glottique) et de sa réponse impulsionnelle (Deller et al. 1993) :

$$x(t) = e(t) * v(t)$$

où $x(t)$ est le signal de parole, $v(t)$ est la réponse impulsionnelle du système modélisant le conduit vocal, et $e(t)$ est le signal d'excitation ayant pour origine les cordes vocales, et $*$ désigne le produit de convolution. La transformée de Fourier du signal transforme la convolution en produit décrit par l'équation suivante

$$X_w(f) = E_w(f) \times V(f)$$

où f dénote la fréquence, $X_w(f)$ *et* $E_w(f)$ sont, respectivement, les spectres d'amplitude des trames d'analyse du signal et de l'excitation glottique pondérées par la fenêtre $w(t)$ et $V(f)$ est la réponse fréquentielle du conduit vocal.

En prenant le logarithme complexe des deux membres de l'équation, il vient :

$$\log(X_w) = log|X_w(f)| + j\angle X_w$$

où $\angle X_w$ dénote la phase de $X_w(f)$. La partie réelle du logarithme complexe représente le spectre d'amplitude du signal et la partie imaginaire représente la phase du signal. La phase doit être corrigée en ajoutant des multiples de$\pm 2\pi$ (Stark et al. 2011).

3.2 Estimation du spectre d'amplitude de la source glottique

Le spectre d'amplitude d'une trame pondérée du signal de parole peut être écrit comme

$$|X_w(f)| = |E_w(f) \times v(f)|$$

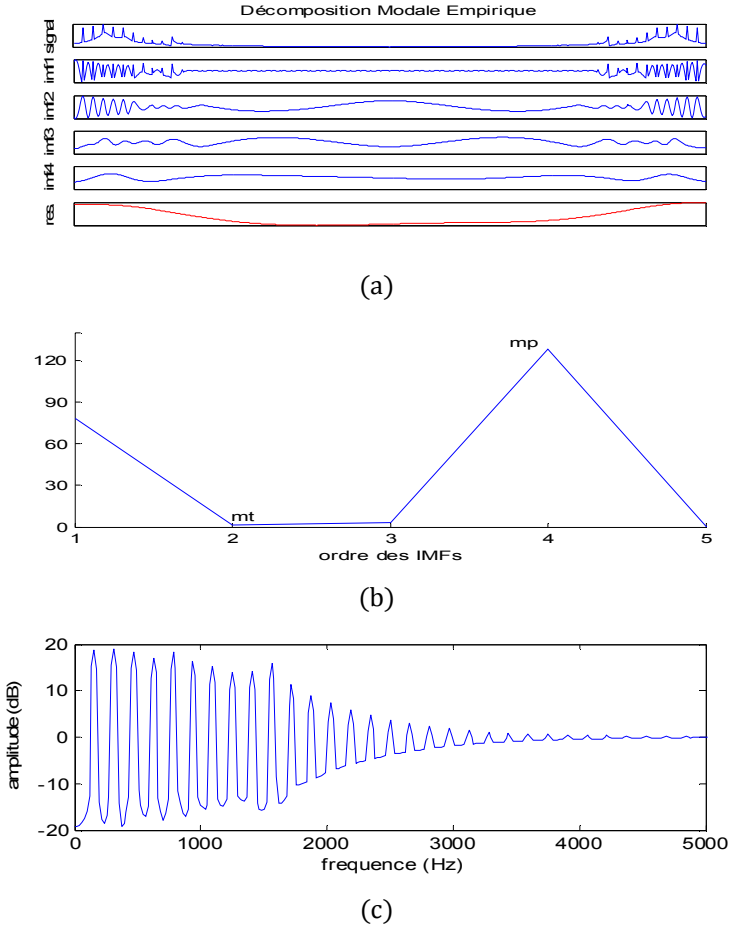
En prenant le logarithme des deux membres de l'équation, il vient :

$$log|X_w(f)| = log|E_w(f)| + \log|v(f)|$$

On observe que le logarithme du spectre d'amplitude d'une trame pondérée du signal de parole est la somme de deux composantes spectrales : le logarithme du spectre d'amplitude de l'excitation pondérée $log|E_w(f)|$, et l'enveloppe spectrale $log|v(f)|$ (Kacha et al. 2012). Le logarithmique du spectre d'amplitude du signal de parole voisée peut être considéré comme constitué d'une variation lente (par rapport à la fréquence) représentant le contour due à la contribution du conduit vocal et d'une série d'harmoniques, caractérisée par une structure périodique. L'algorithme de décomposition modale empirique donne un outil efficace qui permet de séparer les deux composantes du spectre d'amplitude. En effet, l'algorithme de décomposition modale empirique agit comme un banc de filtres (Flandrin et

al. 2004), de sorte que la décomposition du logarithme du spectre d'amplitude résulte en plusieurs composantes oscillantes (IMF) qui peuvent être regroupés en deux catégories (classes) où chaque classe de composants est associée à une partie du spectre d'amplitude.

Il a été montré que la variance des IMFs pour les signaux de parole diminue de manière significative après la quatrième IMF au fur et à mesure que l'ordre des IMFs augmente (Chatlani et al. 2012). Il a été constaté expérimentalement que pour le signal de parole, les statistiques des IMFs sont caractérisées par un pic d'énergie à une IMF d'ordre élevé. Cette propriété est utilisée pour sélectionner l'indice optimal qui permet de séparer la composante harmonique et l'enveloppe spectrale. Les différentes étapes de la méthode de séparation de la composante harmonique et de l'enveloppe spectrale illustrées par la figure 1 peuvent être résumées comme suit (Chatlani et al. 2012) :



(a)

(b)

(c)

FIGURE 1 – Illustration de l'algorithme d'estimation du logarithme du spectre d'amplitude de la source glottique: (a) Décomposition modale empirique du logarithme du spectre d'amplitude d'une trame extraite d'une voyelle /a/ synthétique. (b) Variance des IMFs. (c) Logarithme du spectre d'amplitude de la source glottique.

1. Calculer la variance V de chaque IMF (Figure 1 (a)).

2. Identifier l'indice du maximum, mp dans V pour ordre de l'IMF supérieur a 4.

3. Identifier l'indice du minima mt.

4. Calculer mb tel que $mb = mp - mt$.

5. Déterminer d'indice M avec $M = mp - mb$.

Le logarithme du spectre d'amplitude de la source glottique (Figure 1 (c)) est estimé par la relation suivante:

$$log|E_w(f)| = \sum_{j=1}^{M-1} IMF_j(f)$$

3.3 Estimation de la source glottique

L'approche d'estimation de la source glottique combine le logarithme du spectre d'amplitude de la source glottique et la phase estimée comme la partie imaginaire du logarithme complexe du spectre du signal de parole. L'approche d'estimation est illustrée par la figure 2. Le filtre passe bas a pour rôle d'éliminer l'effet de la fenêtre de pondération $w(t)$.

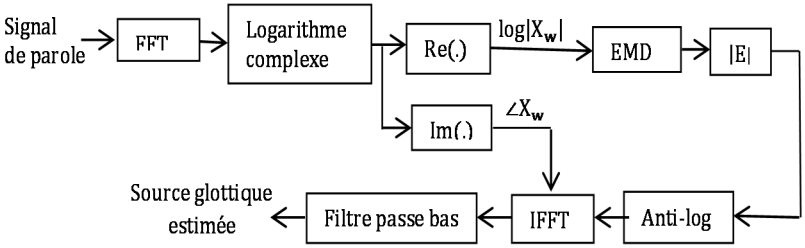
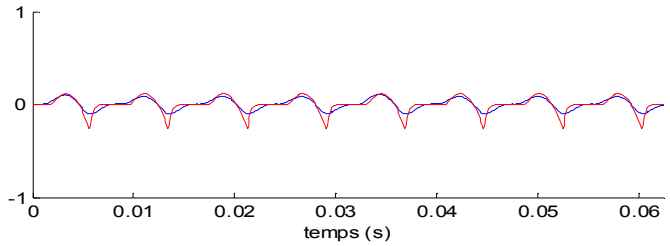


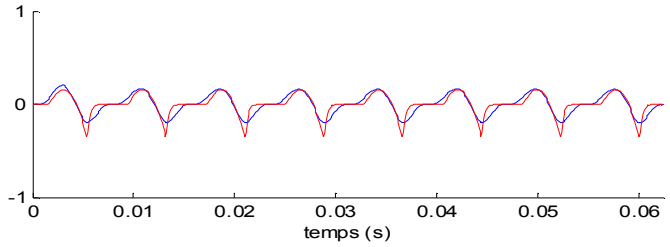
FIGURE 2 – Estimation de la source glottique.

4 Résultats et discussions

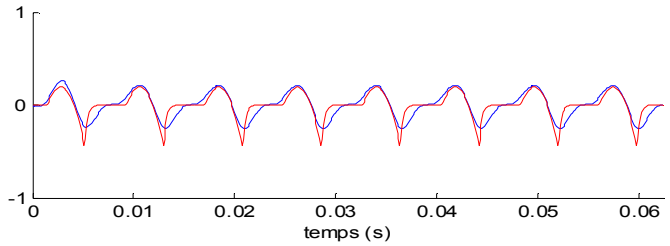
Évaluer objectivement et quantitativement une méthode d'estimation du signal glottique exige de travailler avec des signaux synthétiques, puisque la véritable source n'est pas disponible pour les signaux de parole réelle. Dans ce travail, le signal artificiel utilisé dans le test est la voyelle synthétique /a/ de durée 1 seconde et de fréquence fondamental f0=128 Hz générée selon le modèle source-filtre de production de la parole. La fréquence d'échantillonnage de l'ensemble des signaux de parole utilisés dans l'expérience est de 20 kHz. Le modèle source-filtre est constitué d'une source qui génère un train d'impulsions périodique modélisant le flux d'air glottique et un conduit vocal modélisé comme un filtre



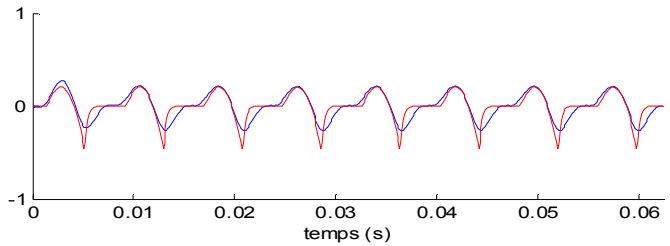
L=256



L=512



L=1024



L=2048

FIGURE 3 – Forme d'onde de la source glottique estimée par la méthode proposée superposée au modèle de l'excitation glottique pour différentes longueurs L de la fenêtre de pondération.

tout-pôle, caractérisé par trois pôles (L. R. Rabiner. 1968; Deller et al. 1993) correspondant aux fréquences de formants 981,6 Hz, 1631,3 Hz et 3165,9 Hz présentant des bandes fréquentielles de 140 Hz, 180 Hz et 55 Hz, respectivement. Le rayonnement aux lèvres est modélisé par un dérivateur du premier ordre $R(z) = 1 - z^{-1}$.

Le signal de parole a été divisé en k trames sans chevauchement en utilisant une fenêtre de Hamming et la source glottique est estimée pour chaque trame. Afin d'étudier les performances de la méthode proposée, l'estimation de la source glottique a été effectuée pour différentes tailles de la fenêtre. Les résultats obtenus sont présentés par la figure 3. La décomposition modale empirique permet d'avoir des estimations précises de la source glottique quelle que soit la longueur de la fenêtre utilisée. La figure 4 montre le modèle de la source glottique utilisée ainsi que les estimations basées sur la décomposition modale empirique et le cepstre complexe pour une trame de longueur 1024 échantillons. Il est observé que l'estimation basée sur la décomposition modale empirique est plus précise que celle basée sur le cepstre complexe.

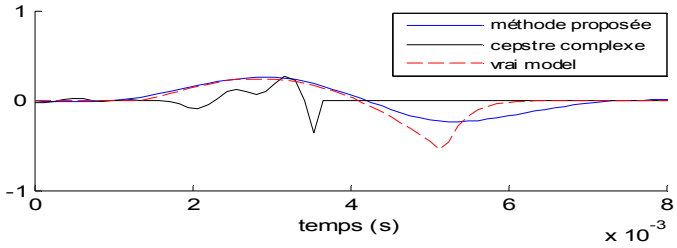


FIGURE 4 –Forme d'onde de la source glottique estimée par la méthode proposée comparée à celle estimée par le cepstre complexe et au vrai modèle pour une trame de longueur L=1024.

5 Conclusion

Dans cette présentation, l'algorithme de décomposition modale empirique a été proposé comme alternative pour estimer l'excitation glottique à partir du signal de parole. La performance de la méthode proposée a été comparée à celle de l'estimation basée sur le cepstre complexe. La méthode proposée est simple et systématique. Les résultats obtenus montrent que la méthode proposée fournit une estimation précise de l'excitation glottique aussi bien pour les trames de courte durée que pour les trames de longue durée.

Références

L. R. Rabiner. (1968). Digital-Formant Synthesizer for Speech-Synthesis Studies. *J. Acoust. Soc. Amer*, 43(4): 822-828.

Alku, J. Svec, E. Vilkman, F. Sram. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3): 109-118.

J. H. Deller, J. G. Proakis, J. H. L. Hansen. (1993). *Discrete-time processing of speech signals*. Prentice-Hall.

D. Brookes, D. Chan. (1994). Speaker characteristics from a glottal airow model using glottal

inverse filtering. *Proc. Institue of Acoust*, 15:501-508.

P. Alku, E. Vilkman. (1994). Estimation of the glottal pulseform based on discrete all-pole modeling. *Third international Conference on Spoken Language Processing*: 1619-1622.

Huang N.E. et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc. R. Soc. London Ser. A* ,454: 903-995.

M. Plumpe, T. Quatieri, D. Reynolds. (1999). Modeling of the glottal flow derivative wave form with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7: 569-586.

D. Yamada, N. Kitaoka, S. Nakagawa. (2002). Speech Recognition Using Features Based on Glottal Sound Source. *Trans. of the Institute of Electrical Engineers of Japan*, 122(12): 2028-2034.

B. Doval, C. dAlessandro, N. Henrich. (2003). The voice source as a causal/anticausal linear filter. *Proceedings ISCA ITRW VOQUAL03*:*15-19*.

E. Moore, M. Clements, J. Peifer, L.Weisser. (2003). Investigating the role of glottal features in classifying clinical depression. *Proc. of the 25th International Conference of the IEEE Engineering in Medicine and Biology Society*, 3:2849-2852.

P. Flandrin, G. Rilling, and P. Goncalves. (2004). Empirical mode decomposition as a filter bank. *IEEE Sig. Proc. Lett*, 11(2): 112–114.

B. Bozkurt, B. Doval, C. DAlessandro, T. Dutoit. (2005). Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech. *IEEE Signal Processing Letters*, 12(4): 2005.

T. Drugman, G. Wilfart, A. Moinet, T. Dutoit. (2009). Using a pitch-synchronous residual for hybrid HMM/frame selection speech synthesis. *IEEE International Conference on Acoustic Speech and Signal* Processing,ICASSP 2009: 3793-3796.

T. Drugman, B. Bozkurt, T. Dutoit, (2009), Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*. Interspeech 2009*: 116-119.

M. Stark, M. Wohlmayr, F. Pernkopf. (2011). Source–Filter-Based Single-Channel Speech Separation Using Pitch Information. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 19(2):242-255.

N. Chatlani, J. Soraghan. (2012). EMD-Based Filtering (EMDF) of Low-Frequency Noise for Speech Enhancement. *IEEE Trans. audio, speech, and lang. proc*,20(4): 1158-1166.

A. Kacha, F. Grenez, J. Schoentgen. (2012). Assessment of Disordered Voices Using Empirical Mode Decomposition in the Log-Spectral Domain. *Interspeech 2012.*

T. Drugman, B. Bozkurt, T. Dutoit. (2012). A comparative study of glottal source estimation techniques. *Computer Speech & Language* ,26(1): 20-34.

A. Kacha, F. Grenez, J. Schoentgen. (2013). Empirical Mode Decomposition-Based Spectral Acoustic Cues for Disordered Voices Analysis, *Interspeech 2013.*