A Long-Term Harmonic plus Noise Model for Speech Signals

Faten Ben Ali[1,2]*, Laurent Girin*[1]*, Sonia Djaziri Larbi*[2]

[1]GIPSA-lab, Grenoble Institute of Technology, Grenoble, France
[2]Unité Signaux et Systèmes, Ecole Nationale d'Ingénieurs de Tunis, Tunisie
{Faten.Ben-Ali, Laurent.Girin}@gipsa-lab.grenoble-inp.fr, sonia.larbi@enit.rnu.tn

Abstract

The harmonic plus noise model (HNM) is widely used for spectral modeling of mixed harmonic/noise speech sounds. In this paper, we present an analysis/synthesis system based on a long-term two-band HNM. "Long-term" means that the time-trajectories of the HNM parameters are modeled using "smooth" (discrete cosine) functions depending on a small set of parameters. The goal is to capture and exploit the long-term correlation of spectral components on time segments of up to several hundreds of ms. The proposed long-term HNM enables joint compact representation of signals (thus a potential for low bit-rate coding) and easy signal transformation (e.g. time stretching) directly from the long-term parameters. Experiments show that it can be compared favourably with the short-term version in terms of parameter rates and signal quality.
Index Terms: speech analysis/synthesis, harmonic + noise model, long-term processing.

1. Introduction

In speech/music coders and analysis/synthesis systems, spectral parameters are usually extracted and processed on a short-term (ST) basis, i.e. every 20ms or so. This is mainly due to the non stationarity of audio signals and/or real-time processing constraints. For speech signals, the evolution of the vocal tract shape and glottal source activity is often quite smooth and regular, and it can be captured in terms of slow AM/FM modulations. High correlation between successive ST spectral parameters is actually exploited for two or three consecutive ST frames in, e.g., differential coding, matrix quantization, or recursive coding [1]. But for non-real-time applications (e.g., half-duplex communication, storage, or transformation), the analysis-synthesis process can be applied on a *long-term* (LT) basis, i.e. much larger signal windows. This is the foundation of the Temporal Decomposition (TD) technique [2], which consists of decomposing the trajectory of spectral parameters into "target vectors" which are sparsely distributed in time and linked by interpolative functions. TD has been recently revisited in [3], where the trajectories of ten consecutive spectral vectors are modeled by fourth-order polynomials.

All those mentioned studies concern Linear Prediction Coding (LPC) parameters, widely used for speech coding. The LT approach can be extended to other spectral models, like sinusoidal models [4][5]. Thus, it was proposed in [6] to model the LT trajectory of harmonic phases and amplitudes with a Discrete Cosine Model (DCM). In [6], LT frames are continuously harmonic sections with very variable size and shape. A fitting algorithm was proposed to automatically adjust the LT model.

The present paper presents a new extension of [6][7] to the framework of the Harmonic + Noise Model (HNM), which is particularly appropriate for modeling mixed voiced/unvoiced speech signals. We focus on a two-band HNM, inspired by [8], and we present the application of the LT modeling approach to the parameters of this two-band HNM. Those parameters are the spectral envelope (of both harmonic and noise amplitudes), the fundamental frequency F_0, and the voicing cut-off frequency F_V. Note that in [6][7], only purely voiced sections were considered. In the present study, we generalize the modeling of the spectral envelope to any harmonic/noise combination, we introduce the LT modeling of F_V, and we simplify the modeling of F_0 (w.r.t. [6]). We also describe and assess a complete original analysis-synthesis system based on the proposed LT-HNM.
This paper is organized as follows. The HNM model is presented in Section 2. The LT modeling of HNM parameters trajectories is described in Section 3, resulting in a complete LTHNM analysis-synthesis system. Experiments and results are given in Section 4.

2. The Two-Band Harmonic + Noise Model

In its general non-stationary form, we can express the HNM as:

$$s(n) = \sum_{h=1}^{H} a_h(n) cos[\phi_h(n)] + \nu(n), \qquad (1)$$

where h is the harmonic rank, H is the number of harmonics, $a_h(n)$ is the harmonic instantaneous amplitude, $\phi_h(n)$ is the instantaneous phase, and $\nu(n)$ is the noise part of the signal. $\phi_h(n)$ is (the sampled version of) the summation of instantaneous frequency over time (each frequency being the h-multiple of F_0). For speech signals, those parameters are assumed to (slowly) vary in time, with possible "birth" and "death" of sinusoids, as in the more general sinusoidal model [4].

In the present work, we use a simplified two-band version of the HNM in the spirit of [8]. This version splits the frequency band into a harmonic band in the low frequency (LF) region, and a noise band in the high frequency (HF) region. The noise band is assumed to model HF random components with spectral coloration but no clear temporal structure. Those two bands are separated by a "boundary frequency" called the voicing cut-off (VCO) frequency F_V. This two-band HNM model is very flexible in the sense that it can be used to represent purely harmonic frames (F_V equal to the Nyquist frequency F_{Nyq}), purely unvoiced frames ($F_V = 0$), or mixed voiced-unvoiced frames.

In the present study, the parameters of the two-band HNM are first extracted on a short-term (ST) basis, as in usual ST-HNM modeling, using analysis frames (indexed by k) of length w = 30ms and hop size r = 20ms. The fundamental frequency F_0 is first estimated using Praat's autocorrelation method [9] (which implicitly provides voiced/unvoiced segmentation). The VCO frequency F_V is estimated using the method of [10] based on the maximization of the sum of a cumulative periodic energy for the lower band and a cumulative aperiodic energy for the

upper band. The estimated F_V value is rounded to the nearest harmonic frequency, which becomes the last harmonic of the frame, indexed by H_k. The estimation of the H_k harmonic amplitude parameters a_h (and initial phases ϕ_h) of the harmonic band is then made by least-square fitting between the (stationary) harmonic model (using the measured F_0) and the signal within the k-th frame [5]. Finally, the frequencies f_n and amplitudes a_n of the N_k spectral components of the noise band are estimated by a peak-picking algorithm [4], applied on the upper band of the FFT magnitude spectrum.

3. LT modeling of HNM parameters

3.1. The LT model and associated estimation process

LT modeling of HNM parameters consists of 1) defining LT frames: in the present study a LT frame is either a continuously voiced (actually mixed voiced/unvoiced) or continuously unvoiced section of speech (as a sequence of K successive ST frames; LT frame boundaries are provided by the F_0 analysis), and 2) Representing the trajectories of the HNM parameters on each LT frame by a sparse P-order time model. The goal is to reduce the data dimension from K to $P+1$, with P being significantly lower than K, while preserving the essential shape of data trajectory. In the present study, we use a linear combination of cosine functions (called Discrete Cosine Model – DCM), since this model has provided good fitting and computational properties in previous studies [6][7]:

$$\tilde{S}_m(n) = \sum_{p=0}^{P} c_{m,p} \cos(p\pi \frac{n}{N}). \quad (2)$$

S_m is either F_0, F_V, or a parameter of the spectral envelope of a_h and a_n (see next section), N denotes the maximum value of data index n. The vector of M spectral parameters extracted at time instant $n_k = kr$ is denoted $\mathbf{S}_k = [S_{1,k} S_{2,k} \dots S_{M,k}]^T$ (^{T} denotes the transpose operator; M is possibly equal to 1 for F_0 or F_V). Thus, we actually have to model M trajectories of K values $S_{m,k}$ which are gathered in the $M \times K$ matrix $\mathbf{S}$. Let us denote by $\mathbf{M}$ the $(P+1) \times K$ "model matrix" of general term $m_{p,k} = \cos(p\pi \frac{n_k}{N})$, and let us denote by $\mathbf{C}$ the $M \times (P+1)$ vector/matrix of model coefficients $c_{m,p}$. When the order P is known, $\mathbf{C}$ is estimated by minimizing the weighted mean square modeling error (WMSE) at the n_k instants, leading to:

$$\mathbf{C} = \mathbf{S} \cdot \mathbf{W} \cdot \mathbf{M}^T \cdot (\mathbf{M} \cdot \mathbf{W} \cdot \mathbf{M}^T)^{-1}, \quad (3)$$

$\mathbf{W}$ is a diagonal weight matrix that can be introduced to control the contribution of the data in the model computation. Also, a diagonal "regularizing" term can be added to the inverted matrix in (3) to fix possible ill-conditioning problems [11]. We do not detail this technical aspect here.

3.2. Model orders optimization

In the previous subsection, P is assumed to be known. In fact, for each LT speech section and each HNM parameter, the goal of efficient LT modeling is to automatically set the model order to a value that ensures a good trade-off between data compression (ideally $P << K$) and good modeling accuracy. For this aim, we propose the following algorithms.

3.2.1. LT modeling of F_0

For F_0 modeling, we define a target ratio $D_t^{F_0}$ for the modeling error (e.g. 1%) and apply the dichotomic search of Algorithm 1. Note that the last iteration is validated only if it leads to lower the error, and since all time frames are here assumed to have the same importance, all weights of W are set to 1. Of course, more refined fitting criteria and strategy can be used, e.g. perceptual criteria with adaptive time-weights [6].

Algorithm 1:

$P \leftarrow$ power of 2 closest to $K/2$, $\Delta P \leftarrow P/2$, and $\mathbf{S} \leftarrow \mathbf{F}_0$
while $\Delta P \geq 1$ **do**
  $\Delta P \leftarrow \Delta P/2$
  Calculate $\mathbf{C}$ with (3)
  Calculate the modeling WMSE E and the relative error $D = E$/mean($\mathbf{F}_0$)
  if $D \leq D_t^{F_0}$ **then**
    $P \leftarrow P - \Delta P$
  else
    $P \leftarrow P + \Delta P$
  end if
end while
$P_{F_0} \leftarrow P$ and $\mathbf{C}_{F_0} \leftarrow \mathbf{C}$

3.2.2. LT modeling of F_V

LT modeling of F_V is similar to the LT modeling of F_0, resulting in optimal $\mathbf{C}_{F_V}$ vector and P_{F_V} order. However, the modeled vectors $\tilde{\mathbf{F}}_\mathbf{v} = \mathbf{C} \cdot \mathbf{M}$ are rounded to the closest harmonic frequency, and the target error $D_t^{F_V}$ is expressed in terms of maximal deviation in (integer) number of harmonics, i.e. P_{F_V} is found as the minimum order so that the maximum modeling error remains within $\pm Q$ harmonic (Q can be set to 1 or 2).

3.2.3. 2D-DCM modeling of the spectral amplitudes

The amplitudes are LT modeled using a 2D modeling approach similar to the one presented in [7] for purely harmonic spectra. This technique is here extended to mixed harmonic/noise sections of speech. The general principle is a two-step modeling: For each ST frame k of a given LT frame, a first DCM model of order M (cf. section 3.1) is applied in the frequency dimension, covering both harmonic and noise amplitudes. This model is similar to the discrete cepstrum proposed in [11][12]. Then a second DCM of order P is applied on the resulting coefficients along the time dimension. M is variable from one LT section to the other but it is the same for all ST frames of the LT section. This enables (i) to switch from a $H_k + N_k$ variable-size set of ST amplitudes to a fixed-size set of parameters that is suitable for LT modeling with (2), and (ii) to reduce the size of the parameter set to be time-modeled, since M is generally significantly lower than $H_k + N_k$. This is a major point for potential coding applications. For the same reason, we also want P to be significantly lower than K, as for F_0 and F_V modeling.

Therefore we propose the two-step Algorithm 2 to find an optimal joint setting for both M and P, ensuring both compact representation and modeling quality. In this algorithm, $\mathbf{M}_k$ is the concatenation of the $H_k \times M$ matrix of general term $m_{h,m} = \cos(m\pi h F_0(k)/F_{nyq})$ and the $N_k \times M$ matrix of general term $m_{n,m} = \cos(m\pi f_n(n,k)/F_{nyq})$. E_{t1} and E_{t2} are user-defined target errors with $E_{t1} < E_{t2}$. The search intervals are set to reasonable values, adapted to speech signals and LT frame length K. The algorithm can be refined with a dichotomic search similar to the one in Algorithm 1 for faster convergence. Also, because of the two-step structure, the algorithm may miss a better (M, P) combination in the area of (M_{opt}, P_{opt}). It can thus be completed with additional search within, e.g., $(M_{opt}+i, P_{opt}-j)$ or $(M_{opt}-i, P_{opt}+j)$ with

$i, j \in [1, 2]$. Finally, **W** is here used in the first part of Algorithm 2 to give more importance to the harmonic amplitudes (weights set to 10) than to the noise peaks (weights set to 1). This was shown to ensure higher global quality for synthesized signals. In future works, a more rigorous criterion will have to be defined and tested regarding this important point.

Algorithm 2:
First part
for $M = M_{min}$ to M_{max} **do**
for $k = 1$ to K **do**
Concatenate harmonic and noise amplitudes into $\mathbf{A}_k = [a_h(1,k), \cdots, a_h(H_k,k), a_n(1,k), \cdots, a_n(N_k,k)]^T$ and calculate the corresponding model matrix $\mathbf{M}_k$.
Calculate the coefficients vector $\mathbf{D}_k$ of the frequency-DCM using a transposed version of (3) applied to $\mathbf{A}_k$ and $\mathbf{M}_k$: $\mathbf{D}_k = (\mathbf{M}_k^T \cdot \mathbf{W} \cdot \mathbf{M}_k)^{-1} \cdot \mathbf{M}_k^T \cdot \mathbf{W} \cdot \mathbf{A}_k$.
Decode the modeled amplitude vectors: $\tilde{\mathbf{A}}_k = \mathbf{M}_k \cdot \mathbf{D}_k$.
end for
Calculate the WMSE E_1 between all original and modeled amplitudes over the whole LT section.
if $E_1 \leq E_{t1}$ **then**
$M_{opt} \leftarrow M$
return (end of Part 1)
end if
end for

Second part
for $P = P_{min}$ to P_{max} **do**
Calculate the $(M_{opt}+1) \times (P+1)$ coefficients matrix **C** of the time-DCM by applying (3) to the spectral envelope matrix $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_K]$.
Decode the modeled spectral envelope matrix $\tilde{\mathbf{D}} = \mathbf{C} \cdot \mathbf{M}$. Let $\tilde{\mathbf{D}}_k$ denotes the k-th column of $\tilde{\mathbf{D}}$.
for $k = 1$ to K **do**
Decode the new modeled amplitudes $\tilde{\mathbf{A}}_k = \mathbf{M}_k \cdot \tilde{\mathbf{D}}_k$.
end for
Calculate the new WMSE E_2 between all original and modeled amplitude values.
if $E_2 \leq E_{t2}$ **then**
$P_{opt} \leftarrow P$ and $\mathbf{C}_{opt} \leftarrow \mathbf{C}$
return
end if
end for

3.3. Synthesis of LT modeled signals

As for the signal synthesis, the harmonic part $\tilde{s}_h(n)$ is obtained in a very straightforward manner: the phase of each h-harmonic is obtained by summation of the h-multiple of the modeled F_0 trajectory provided by (2) (plus the initial phase of the first ST frame of the LT section to preserve natural sounding). The modeled harmonic amplitudes are linearly interpolated between ST frames and $\tilde{s}_h(n)$ is obtained by components summation as in (1). The noise part $\tilde{\nu}(n)$ is generated with the overlap-add random phase sinusoid technique of [13], using amplitudes sampled from the modeled spectral envelope in the noise band every $\Delta f = 70$ Hz. Note that the modeled F_0 and F_V trajectories are independent, thus the upper harmonic trajectories may be "interrupted" by noise regions. This is managed using a "birth and death" process (involving local interpolation of a_h to 0) [4].

4. Experimental results

4.1. Data

We report the results of the LT modeling of 24 speech sentences sampled at 16kHz (13 male and female speakers and 4 languages), of total duration 50s. LT segmentation resulted in 291 LT sections with a mean duration of 0.17s and a maximum duration of 1.24s [136 voiced sections (33s) and 155 unvoiced sections (17s)]. The ST analysis hop size is r =20ms.

4.2. Compression gain of the LT modeling

We provide here the coefficient rates for the ST-HNM parameters set and the LT-HNM model, respectively denoted R_{ST} and R_{LT}. We also compare the LT-HNM with a ST version with DCM modeling of the spectral envelope only, denoted 1D-HNM. For fair comparison, the envelope model of this 1D-HNM is calculated with a target error equal to E_{t2}^{LT}. The coefficient rate of the 1D-HNM is denoted R_{1D} and we have, for a given LT section, $R_{ST} = [\sum_{k=1}^{K}(2 + 2(H_k + N_k))]/(K.r)$, $R_{1D} = [K(M_{opt} + 3) + \sum_{k=1}^{K} H_k]/(K.r)$ and $R_{LT} = [(M_{opt}+1)(P_{opt}+1)+P_{F_0}+P_{F_v}+H_0+3]/(K.r)$, where H_0 is the number of initial phases (issued from the first ST frame).

For the used database, the mean rate is $R_{ST} = 6298$ coeff/s. For example, when applying the LT-HNM with target errors $(E_{t1}^{LT}, E_{t2}^{LT}) = (0.6, 0.7)$dB, we obtain $R_{LT} = 530$ coeff/s, while $R_{1D} = 983$ coeff/s (with 1D-target error $E_{t1}^{1D} = 0.7$dB). Hence, the LT-modeling achieves a rate gain of 91.5% compared to the ST-HNM[1] and 46% compared to the 1D-HNM. To evaluate the compression gain due to 2D-DCM amplitude modeling only, we provide in Table1 the corresponding coefficients rates R_{1D}^A and R_{LT}^A for different target error settings. It can be seen that the coefficient rates decrease when the target errors increase (as expected) and that, for a given E_{t2}^{LT}, the best LT combination is systematically the one with E_{t1}^{LT} immediately higher (i.e. the lower diagonal). This suggests that a more efficient overall LT modeling is obtained when the modeling of the spectral envelope is not much constrained. When comparing the optimal R_{LT}^A with R_{1D}^A, we observe important rate gains that increase with target errors (up to 38%).

To illustrate the efficiency of the proposed LT-HNM, we plot in Fig. 1 the DCM orders P_{F_0} and P_{F_V} as a function of the voiced LT sections length (number of ST frames K). An average gain of about 2 is obtained. The 2D-amplitude modeling

[1]Here the gain is not only due to the LT-modeling of F_0, F_v and the spectral amplitudes, but also because the noise frequencies f_n and the harmonic phases ϕ_h do not need to be sent to the decoder.

E_{t1}^{LT} \ E_{t2}^{LT}	0.6	0.7	0.8	0.9	1.0	1.1
0.4	858	738	648	585	544	496
0.5	648	531	466	418	383	351
0.6	-	433	367	324	296	269
0.7	-	-	315	274	246	225
0.8	-	-	-	246	217	197
0.9	-	-	-	-	202	181
1.0	-	-	-	-	-	173
R_{1D}^A	668	512	420	358	315	284
Gain (%)	2	15	25	31	35	38

Table 1: Coefficient rate R_{LT}^A of 2D-DCM amplitude modeling for different target errors (E_{t1}^{LT}, E_{t2}^{LT}), and coefficient rate R_{1D}^A of the reference 1D-HNM with E_{t1}^{1D}=E_{t2}^{LT}, $R_{ST} = 6289$.

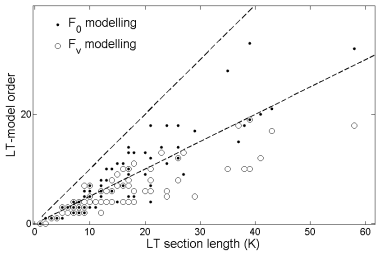

Figure 1: *LT model order of F_0 (.) and F_V (o) trajectories as a function of LT section length K for all test sections. The dashed lines are the 1^{st} and 2^{nd} bisectors. $D_t^{F_0} = 1\%$, $D_t^{F_V} = 8\%$.*

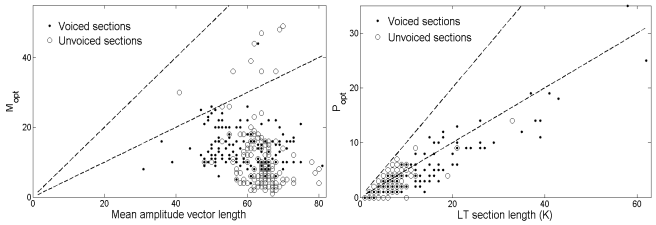

Figure 2: *2D amplitude modeling: M_{opt} as a function of ST amplitude vectors length (left) and P_{opt} as a function of LT frame length K (right). $(E_{t1}^{LT}, E_{t2}^{LT}) = (0.6, 0.7)dB$.*

gain is illustrated on Fig. 2 which depicts the LT model orders M_{opt} and P_{opt} as a function of the mean value of $(H_k + N_k)$ over the LT section and K, respectively. The average coefficients gain is also around 2 for time-modeling, while it is more important (around 4) for frequency modeling, although with a significantly scattering. Fig. 3 provides an example of original and modeled trajectories of F_0 and of the first harmonic amplitude. Both plots show that the modeled trajectories follow the original ones quite fairly given the low number of coefficients.

4.3. Quality of synthesis signals

The perceptual quality of the speech test signals modeled with the LT-HNM [for $(E_{t1}^{LT}, E_{t2}^{LT}, D_t^{F_0}, D_t^{F_v}) = (0.6, 0.7, 1\%, 8\%)$] and the reference 1D-HNM [for $E_{t1}^{1D} = 0.7dB$] was assessed with PESQ[2] . The obtained mean scores are 2.4 and 2.3 respectively, while the ST-HNM score is 2.9. These results indicate that the LT-HNM model provides nearly the same signal quality than the 1D-HNM, while significantly reducing the coefficients rate.

5. Conclusion-Perspectives

A "flexible" LT-HNM model was presented, using DCM models for F_0, F_V and spectral amplitudes. Compared to the ST and 1D versions, a significant gain in coefficients rate was obtained. The proposed algorithms for setting the DCM orders enable a trade-off between coefficients rate gain and modeling quality. Future works will concern i) the use of perceptual criteria for LT model fitting, ii) a better LT modeling of the phase trajectories, iii) the use of the proposed model for speech transformation (e.g. time-stretching) and iv) the design of a low-bitrate LT speech coder based on the proposed LT-HNM, with "LT-quantization" of HNM parameters in the line of the one proposed in [14] for LPC parameters.

[2]Perceptual Evaluation of Speech Quality, ITU-T Recom. P.862.

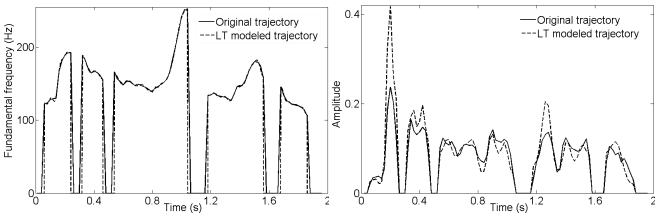

Figure 3: *Trajectories of LT-modeled F_0 (left) and of the 1^{st} 2D-modeled harmonic amplitude (right) $(E_{t1}^{LT}, E_{t2}^{LT}) = (0.6, 0.7)dB$.*

6. References

[1] R. M. Gray and A. Gersho, Vector Quantization and Signal Compression, Kluwer Acad. Pub., Boston, Mass, 1992.

[2] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," Proc. IEEE ICASSP, Boston, 1983.

[3] S. Dusan, J. Flanagan, A. Karve and M. Balaraman, "Speech compression by polynomial approximation," IEEE Trans. Audio, Speech, Lang. Proc., 15(2):387-395, 2007.

[4] R. J. McAulay and T. F. Quatieri, "Speech analysis synthesis based on a sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Proc., 34(4), 1986.

[5] E. B. George and M. J. Smith, "Speech Analysis synthesis and modification using an analysis by synthesis overlap add sinusoidal model," IEEE Trans. Acoust., Speech, Signal Proc., 5(5), 1997.

[6] L. Girin, M. Firouzmand and S. Marchand, "Perceptual long term variable rate sinusoidal modeling of speech," IEEE Trans. Speech and Audio Proc., 15(3):851-861, 2007.

[7] M. Firouzmand and L. Girin, "Long-term flexible 2D cepstral modeling of speech spectral amplitude," Proc. IEEE ICASSP, Las Vegas, Nevada, 2008.

[8] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Trans. Acoust., Speech, Signal Proc., 9(1), 2001.

[9] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proc. of the Institute of Phonetic Sciences 17: 97-110. Univ. of Amsterdam, 1993.

[10] K. Hermus, L. Girin, H. Van Hame and S. Irhimeh, "Estimation of the voicing cut-off frequency contour of natural speech based on harmonic and aperiodic energies," Proc. IEEE ICASSP, Las Vegas, Nevada, 2008.

[11] O. Cappé, J. Laroche and E. Moulines, "Regularized estimation of cepstrum enveloppe from discrete frequency points," Proc. IEEE WASPAA, New Paltz, NY, 1995.

[12] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete cepstra: Application to musical sound signals," Proc. Int. Computer Music Conf., Glasgow, UK, 1990

[13] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced Speech," IEEE Trans. Speech and Audio Proc., 5(6):557- 560, 1997.

[14] L. Girin, "Adaptive long term coding of LSF parameters trajectories for large-delay/very-to utra-low bit rate speech coding," Eurasip J. Audio, Speech, Music Proc., 2010, Article ID 597036.