A long-term harmonic plus noise model for speech signals
Faten Ben Ali, Laurent Girin, Sonia Djaziri-Larbi

To cite this version:

HAL Id: hal-00695752
https://hal.archives-ouvertes.fr/hal-00695752
Submitted on 9 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Long-Term Harmonic plus Noise Model for Speech Signals

Faten Ben Ali¹ ², Laurent Girin¹, Sonia Djaziri Larbi²

¹GIPS-lab, Grenoble Institute of Technology, Grenoble, France
²Unité 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] the LT frames are continuously harmonic sections with very variable size and ”shape”, and a fitting algorithm was proposed to automatically adjust the LT model according to the speech section content. In [7], this adaptive scheme was extended to the LT modeling of harmonic amplitudes envelope parameters.

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 LT-HNM analysis-synthesis system. Experiments and results are given in Section 4.

2. The Two-Band Harmonic + Noise Model
In its general non-stationary form, the HNM can be formally expressed as:

$$s(n) = \sum_{h=1}^{H} a_h(n) \cos[\phi_h(n)] + \nu(n),$$  \hspace{1cm} (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 $F_{N\text{q}}$), 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 $w = 30$ms and hop size $r = 20$ms. The fundamental frequency $F_0$ is first estimated using Praat’s autocorrelation method [9]
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]:

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

\(S_m\) is either \(F_0\), \(F_V\), or a parameter of the spectral envelope of \(a_n\) and \(a_0\) (see next section). The vector of \(M\) spectral parameters extracted at time instant \(n_k = kr\) is denoted \(S_k = [S_{1,k}, S_{2,k}, \ldots, S_{M,k}]^{T}\) (\(T\) 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 \(S\). Let us denote by \(M\) the \((P + 1) \times K\) “model matrix” of general term \(m_{p,k} = \cos(p \frac{n}{\Delta n})\), and let us denote by \(C\) the \(M \times (P + 1)\) vector/matrix of model coefficients \(c_{m,p}\). When the order \(P\) is known, \(C\) is estimated by minimizing the weighted mean square modeling error (WMSE) at the \(n_k\) instants, leading to:

\[
C = S \cdot W \cdot M^{T} \cdot (M \cdot W \cdot M^{T})^{-1},
\] (3)

\(W\) is a diagonal weight matrix that can be introduced to control the contribution of the model 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 \(D^{F_0}\) for the modeling error (e.g. 1%) and apply the dichotomous 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:

1. \(P \leftarrow \text{power of 2 closest to } K/2, \Delta P \leftarrow P/2, \text{and } S \leftarrow F_0\)

   while \(\Delta P \geq 1\)

   1. Calculate \(C\) with (3)

   2. Calculate the modeling WMSE \(E\) and the relative error

      \(D = E/\text{mean}(F_0)\)

   if \(D \leq D^{F_0}\) then

      \(P \leftarrow P - \Delta P\)

   else

      \(P \leftarrow P + \Delta P\)

   end if

end while

\(P_{F_0} \leftarrow P\) and \(C_{F_0} \leftarrow C\)

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 \(C_{F_V}\) vector and \(P_{F_V}\) order. However, the modeled vectors \(F_V = C \cdot M\) are rounded to the closest harmonic frequency, and the target error \(D^{F_V}\) 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 (harmonic and noise) 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 \(M\) 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, \(M_k\) is the concatenation of \(H_k \times M\) matrix of general term \(m_{n,m} = \cos(m \pi h_k(n)/F_0)\) and the \(N_k \times M\) matrix of general term \(n_{n,m} = \cos(m \pi f_k(n)/F_0)\). \(E_{11}\) and \(E_{12}\) are user-defined target errors with \(E_{11} < E_{12}\). The search intervals are set to reasonable values, adapted to speech signals and LT frame length \(K\). The algorithm can be refined with a dichotomous search similar to the one in Algorithm 1 for faster convergence. Also, because of the two-step structure, the al-
Algorithm may miss a better \((M, P)\) combination in the area of \((M_{\text{opt}}, P_{\text{opt}})\). It can thus be completed with additional search within, e.g., \((M_{\text{opt}} + i, P_{\text{opt}} - j)\) or \((M_{\text{opt}} - i, P_{\text{opt}} + j)\) 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 \(M = M_{\text{min}}\) to \(M_{\text{max}}\) do  
for \(k = 1\) to \(K\) do  
Concentrate harmonic and noise amplitudes into \(A_k = [a_1(k), a_2(k), \ldots, a_n(k)]^T\) and calculate the corresponding model matrix \(M_k\).  
Decompute the coefficients vector \(D_k\) of the frequency-DCM using a transposed version of \(D\) applied to \(A_k\) and \(M_k\):  
end for  
end for  
Second part  
for \(P = P_{\text{min}}\) to \(P_{\text{max}}\) do  
Calculate the coefficients matrix \(C\) of the time-DCM by applying \((P + 1)\) to the spectral envelope matrix \(D = [D_1, D_2, \ldots, D_K]\).  
Decompute the modeled spectral envelope matrix \(\tilde{D} = C \cdot M\).  
end for  
end for

4. Experimental results

4.1. Data

We report the results of the LT modeling of 24 speech sentences sampled at 16 kHz (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 \(r = 20\)ms.

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 \(E_{1D}\). The coefficient rate of the 1D-HNM is denoted \(R_{1D}\) and calculated for a given LT section, \(R_{LT} = \frac{\sum_{k=1}^{K} (2+2H_k+\bar{N}_k)}{(K,r)}\). \(R_{1D} = \frac{(M_{\text{opt}}+3)\sum_{k=1}^{K} H_k}{(K,r)}\) and \(R_{LT} = \frac{(M_{\text{opt}+1})+P_{\text{opt}}}{P_{\text{opt}}+1}+R_{0}+H_{\text{opt}+1}/(K,r)\), where \(H_{\text{opt}}\) is the number of initial phases (issued from the first ST frame).

For the used database, the mean rate is \(R_{LT} = 6898\) coeff/s. For example, when applying the LT-HNM with target errors \((E_{1D}, E_{1D}^L) = (0.6, 0.7)\)dB, we obtain \(R_{LT} = 530\) coeff/s, while \(R_{1D} = 983\) coeff/s (with 1D-target error \(E_{1D}^L = 0.7\)dB). Hence, the LT-modeling achieves a rate gain of 91.5% compared to the ST-HNM\(^4\) and 46% compared to the 1D-HNM.

To evaluate the compression gain due to 2D-DCM amplitude modeling only, we provide in Table 1 the corresponding coefficient rates \(R_{1D}^2\) and \(R_{LT}^2\) 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 \(E_{1D}^L\), the best LT combination is systematically the one with \(E_{1D}^L\) 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 \(R_{LT}^2\) with \(R_{1D}^2\), 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 \(P_{F_1}\), 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

\[E_{1D}^L = E_{1D}^L + E_{1D}^L\]

Table 1: Coefficient rate \(R_{LT}^2\) of 2D-DCM amplitude modeling for different target errors \((E_{1D}^L, E_{1D}^L)\), and coefficient rate \(R_{1D}^3\) of the reference 1D-HNM with \(E_{1D}^L = E_{1D}^L\).

\[\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
E_{1D}^L / E_{1D}^L & 0.6 & 0.7 & 0.8 & 0.9 & 1.0 & 1.1 \\
\hline
0.4 & 858 & 738 & 648 & 585 & 544 & 496 \\
0.5 & 648 & 531 & 466 & 418 & 383 & 351 \\
0.6 & - & 333 & 367 & 324 & 296 & 269 \\
0.7 & - & - & 274 & 246 & 225 & - \\
0.8 & - & - & - & 217 & 197 & - \\
0.9 & - & - & - & - & 202 & 181 \\
1.0 & - & - & - & - & - & 173 \\
\hline
\end{array}\]

\[\begin{array}{|c|c|c|c|c|c|c|}
\hline
R_{1D}^3 & 668 & 512 & 420 & 358 & 315 & 284 \\
\hline
Gain (%) & 2 & 15 & 25 & 31 & 35 & 38 \\
\hline
\end{array}\]

\(^4\)Here the gain is not only due to the LT-modeling of \(F_0\) and \(F_1\) and the spectral amplitudes, but also because the noise frequencies \(f_n\) and the harmonic phases \(\phi_n\) do not need to be sent to the decoder.
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 \((E^{LT}_{1}, E^{LT}_{2}, D^{F_0}, D^{F_v}) = (0.6, 0.7)dB)\) and the reference 1D-HNM \(E^{LP}_{1} = 0.7dB)\) was assessed with PESQ\(^{\s1}\). 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.

\(^{\s1}\)Perceptual Evaluation of Speech Quality, ITU-T Recom. P.862.

6. References