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

To cite this version:
A Long-Term Harmonic plus Noise Model for Speech Signals

Faten Ben Ali\textsuperscript{1,2}, Laurent Girin\textsuperscript{1}, Sonia Djaziri Larbi\textsuperscript{2}

\textsuperscript{1}GIPSA-lab, Grenoble Institute of Technology, Grenoble, France
\textsuperscript{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. 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 where 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:

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

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 \( v(n) \) 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 \( F_{N_{Nyq}} \)), 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 \( w = 30 \)ms and hop size \( r = 20 \)ms. 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 $H_{0}$. The estimation of the $H_{0}$ harmonic amplitude parameters $a_{n}$ (and initial phases $\phi_{n}$) 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 $N_{H}$ 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 $M$-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]:

$$S_{m}(n) = \sum_{p=0}^{P} c_{m,p} \cos(p \pi n/N). \tag{2}$$

$S_{m}$ 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 $m_{p,k} = \cos(p \pi k/N)$, 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:

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

$W$ is a diagonal weight matrix that can be introduced to control the contribution of 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 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:

1. $P \leftarrow$ power of 2 closest to $K/2$, $\Delta P \leftarrow P/2$, and $S \leftarrow F_{0}$
2. while $\Delta P \geq 1$
   a. Calculate $C$ with (3)
   b. Calculate the modeling WMSE $E$ and the relative error
   $$D = E/\text{mean}(F_{0})$$
   c. if $D \leq D F_{0}$ then $P \leftarrow P - \Delta P$
   else $P \leftarrow P + \Delta P$
3. end

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, $M_{h}$ is the concatenation of the $H_{k} \times M$ matrix of general term $m_{h:m} = \cos(m \pi h F_{0}(k)/F_{0,yq})$ and the $N_{k} \times M$ matrix of general term $m_{n:m} = \cos(m \pi f_{n}(k)/F_{0,yq})$. $E_{11}$ and $E_{12}$ are user-defined target errors $E_{11} < E_{12}$. The search intervals are set to reasonable values, adapted to speech signals and LT frame length $K$. The algorithm can be refined with a dichotomic 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:**

1. **First part**
   - For \(M = M_{\text{min}}\) to \(M_{\text{max}}\)
     - For \(k = 1\) to \(K\)
       - Concatenate harmonic and noise amplitudes into \(\mathbf{A}_k = [\alpha_k(1, k), \ldots, \alpha_k(H_k, k), \alpha_n(1, k), \ldots, \alpha_n(N_k, k)]^T\) 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: \(\hat{\mathbf{A}}_k = \mathbf{M}_k \cdot \mathbf{D}_k\).
     - End for
   - Calculate the WMSE \(E_1\) between all original and modeled amplitudes over the whole LT section.
     - If \(E_1 \leq E_{11}\) then
       - \(M_{\text{opt}} \leftarrow M\)
     - End if
   - End for

2. **Second part**
   - For \(P = P_{\text{min}}\) to \(P_{\text{max}}\)
     - Calculate the \((M_{\text{opt}} + 1) \times (P + 1)\) coefficients matrix \(\mathbf{C}\) of the time-DCM by applying (3) to the spectral envelope matrix \(\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K]\).
     - Decode the spectral envelope matrix \(\hat{\mathbf{D}} = \mathbf{C} \cdot \mathbf{M}\).
     - Let \(\mathbf{D}_k\) denotes the \(k\)-th column of \(\mathbf{D}\).
     - For \(k = 1\) to \(K\)
       - Decode the new modeled amplitudes \(\hat{\mathbf{A}}_k = \mathbf{M}_k \cdot \mathbf{D}_k\).
     - End for
   - Calculate the new WMSE \(E_2\) between all original and modeled amplitude values.
     - If \(E_2 \leq E_{12}\) then
       - \(P_{\text{opt}} \leftarrow P\) and \(C_{\text{opt}} \leftarrow C\)
     - End if
   - End for

3.3. Synthesis of LT modeled signals

As for the signal synthesis, the harmonic part \(\tilde{s}(n)\) 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 \(\tilde{s}(n)\) is obtained by components summation as in (1). The noise part \(\tilde{v}(n)\) 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 \(\alpha_k\) 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 \(r = 20\)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 \(E_{1T}^T\). The coefficient rate of the 1D-HNM is denoted \(R_{LT}^T\) and we have, for a given LT section, \(R_{ST} = \frac{\sum_{k=1}^{K} (2 + 2H_k + N_k))/K.r}{(K.r)}\), \(R_{LT} = \frac{(M_{\text{opt}} + 1)(P_{\text{opt}} + 1) + P_{F_0} + P_{F_v} + H_0 + 3)/K.r}{(K.r)}\) and \(H_0\) is the number of initial phases (issued from the first ST frame).

For the used database, the mean rate is \(R_{LT} = 6298\) coeff/s. For example, when applying the LT-HNM with target errors \((E_{1T}^T, E_{LT}^T) = (0.6, 0.7)\)dB, we obtain \(R_{LT} = 530\) coeff/s, while \(R_{LT} = 983\) coeff/s (with 1D-target error \(E_{1D}^T = 0.7\)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 Table 1 the corresponding coefficient rates \(R_{LT}^T\) and \(R_{LT}^T\) 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 \(E_{1T}^T\), the best LT combination is systematically the one with \(E_{LT}^T\) 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 \(R_{LT}^T\) with \(R_{LT}^T\), 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 \(F_{0T}\) and \(F_{vT}\) 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

\[\begin{array}{c|cccccc}
E_{1T}^T / E_{LT}^T & 0.6 & 0.7 & 0.8 & 0.9 & 1.0 & 1.1 \\
\hline
0.4 & 858 & 738 & 648 & 585 & 544 & 496 \\
0.5 & 648 & 531 & 466 & 418 & 383 & 351 \\
0.6 & 513 & 367 & 324 & 296 & 269 & 245 \\
0.7 & 384 & 294 & 246 & 225 & 206 & 185 \\
0.8 & - & - & - & - & - & - \\
0.9 & - & - & - & - & - & - \\
1.0 & - & - & - & - & - & - \\
\end{array}\]

\[R_{LT}^T\]

\[\begin{array}{c|cccc}
0.6 & 512 & 320 & 288 & 256 \\
0.7 & 420 & 358 & 315 & 284 \\
\end{array}\]

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

\[\text{Gain} (%) = 2 \times 15 / 25 = 31 / 35 = 38\]

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

References