Perceptual Long-Term Variable-Rate Sinusoidal Modeling of Speech

Laurent Girin, *Member, IEEE*, Mohammad Firouzmand, and Sylvain Marchand

***Abstract*—In this paper, the problem of modeling the time-trajectory of the sinusoidal components of voiced speech signals is addressed. A new global approach is presented: a single so-called long-term (LT) model, based on discrete cosine functions, is used to model the overall trajectories of amplitude and phase parameters, for each entire voiced section of speech, differing from usual (short-term) models defined on a frame-by-frame basis. The complete analysis-modeling-synthesis process is presented, including an iterative algorithm for optimal fitting between LT model and measures. A major issue of this paper concerns the use of perceptual criteria in the LT model fitting process (both for amplitude and phase modeling). The adaptation of perceptual criteria usually defined in the short-term and/or stationary cases to the long-term processing is proposed. Experiments dealing with the ten first harmonics of voiced signals show that the proposed approach provides an efficient variable-rate representation of voiced speech signals. Promising results are given in terms of modeling accuracy, synthesis quality, and data compression. The interest of the presented approach for speech coding and speech watermarking is discussed.**

***Index Terms*—Perceptual models, sinusoidal model, speech modeling, speech processing, variable rate.**

I. Introduction

ADDITIVE synthesis [1] is the original spectrum modeling technique. It derives from Helmholtz's research and is rooted in Fourier's theorem, which states that any periodic function can be modeled as a sum of sinusoids at various amplitudes and harmonic frequencies. The most famous technique for additive synthesis is probably the phase vocoder [2]–[4], which is mainly an implementation of the short-time Fourier transform [5]. Also, the experiments conducted by Helmholtz found new developments with computer music, when Risset and Mathews measured the time-varying spectra of several musical instruments [6]. For quasi-stationary pseudoperiodic sounds, the amplitudes and frequencies of the sinusoids continuously evolve slowly with time, controlling a set of pseudo-sinusoidal oscillators commonly called *partials*. Using this model, so-called *sinusoidal model of speech* by McAulay and Quatieri when applied to speech signals in [7], the signal $s(n)$ is given by

$$s(n) = \sum_{i=1}^{I} A_i(n) \cos[\theta_i(n)] \tag{1}$$

with

$$\theta_i(n) = \sum_{k=0}^{n} \omega_i(k) + \theta_i(0) \tag{2}$$

where I is the number of partials, and the functions $A_i(n)$, $\theta_i(n)$, and $\omega_i(n)$ are the instantaneous amplitudes, phases, and digital frequencies (expressed in radians per sample) of the ith partial, respectively, and are slowly evolving with time.

An analysis-synthesis system based on such a model usually requires the measurement of these parameters at the centers of consecutive (possibly overlapping) signal frames, and the interpolation of the consecutive measured values at each sample index to reconstruct the signal by applying (1) and (2) on the interpolated values. For this aim, a key technique called *partial tracking* consists in following the evolutions of power spectrum maxima in time to form the partial trajectories [7], [8]. This model is now used in many software packages like AudioSculpt [9], Lemur [10], PARSHL [11], SMS [12], or InSpect [13]. For real-time additive synthesis, several methods have been proposed, either using the inverse Fourier transform [14] or digital resonators [15], [16]. Also, this approach has been successfully applied to a wide range of applications, such as coding [17] or time- and frequency-stretching [18], [19].

Since the process attempts to follow the time dynamics of speech or music, measures are generally spaced with a delay of a few milliseconds. In the remainder of the paper, we denote by the label short-term (ST) models, the interpolation process that reconstructs the signal samples between two consecutive measures spaced by such ranges of intervals (also, this segment is called the short-term synthesis frame). In such approach, amplitudes are generally interpolated linearly between frames, though more sophisticated methods can be applied (e.g., cubic polynomial interpolation). For frequency/phase trajectories, the problem is slightly more complicated, since the frequencies are the time-derivatives of the phases, and the phase measures are generally provided modulo 2π. Early works considered piecewise linear interpolation of frequencies (e.g., [20]). However, if one wants to accurately reconstruct the signal waveform shape (the *shape-invariant* property of [18]), in addition to measured frequencies, measured phases must be respected, e.g., as with the cubic polynomial interpolation proposed in [7]. Other models were proposed in the literature, e.g., [21], [22].

In this paper, we focus on speech signals, and we propose a different approach to reconstruct the signal from the measured parameters of the sinusoidal model. Instead of interpolating these measures from one analysis frame center to the next, we propose to consider longer sections of speech, beyond the phoneme or syllable level, and model the entire trajectory of the parameters of each partial over the considered section with a single so-called long-term (LT) model (as opposed to the previ-

Manuscript received January 31, 2005; revised June 12, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

L. Girin and M. Firouzmand are with the Speech Communication Laboratory, National Polytechnical Institute of Grenoble, 38000 Grenoble, France (e-mail: girin@icp.inpg.fr; firouz@icp.inpg.fr).

S. Marchand is with LaBRI, Computer Science Laboratory, University of Bordeaux 1, 33405 Talence, France (e-mail: sylvain.marchand@labri.fr).

Digital Object Identifier 10.1109/TASL.2006.885928

ously mentioned short-term models). In this paper, these long sections of speech are continuously voiced sections. In other words, speech is first segmented into voiced and unvoiced parts, then the sinusoidal model is applied to each one of the voiced sections, and a single LT model is used to represent the whole trajectory of either the amplitude or phase of a partial over this section. In this paper, the proposed LT model for amplitude and phase trajectories is based on a linear combination of cosine functions, quite similar to the well-known discrete cosine transform (DCT). Of course, other models are possible, as discussed at the end of this paper.

Several remarks must be made here to better understand this new approach. First, entirely voiced sections generally appear as weakly nonstationary quasi-periodic signals. One can thus simplify the whole analysis-modeling-synthesis process by using the harmonicity assumption: partials become regularly spaced harmonics. Thus, parameter measures of the same rank can be directly associated before applying our LT model on these measures, and the partial tracking is actually avoided.[1] Unvoiced sections are not considered in this paper, since they cannot be modeled efficiently by the sinusoidal/ harmonic model (at least, the way it is processed in this study). Other adequate models can be used for these sections, see, e.g., [8], [11], and [23]–[25], and part of our future work deals with testing adequate LT models for unvoiced speech.

Second, the length and shape of entirely voiced sections can vary widely. For example, it can contain several phonemes or syllables (it can even be a quite long sentence in some cases). That is why we also propose in this paper a method to automatically adjust the complexity (i.e., the order) of the LT model according to the characteristics of the modeled speech section, resulting in a variable-rate modeling process. A generic fitting algorithm, based on an iterative weighted mean square error minimization, is proposed, and adapted for amplitude and phase trajectories modeling. A major point is that this method is based on perceptual optimization criteria: the model order is iteratively adjusted so that a psychoacoustically based criterion is achieved. This criterion is based on a frequency-domain amplitude masking threshold model in the case of amplitude modeling, and it is based on a frequency modulation threshold model in the case of phase modeling. In both cases, the proposed method jointly ensures optimal estimation of the model order and optimal perceptual fitting of the model with the data.

Finally, it can be noted that the idea of modeling the long-run trajectory of speech features was also recently proposed in [26], simultaneously with a preliminary version of our work that only considered phase LT modeling without perceptual criteria [27]. The authors of [26] proposed to model the trajectories of line spectral frequencies (LSFs) parameters by a polynomial, and they implemented a very low bit rate speech coder exploiting this idea. However, this study significantly differs from ours on at least three points: First, we consider sinusoidal parameters instead of LSF. Second, we use an LT model based on discrete cosine functions, whereas in [26], a polynomial model is considered. Third, we deal with variable-order models to capture large variable-size trajectories, whereas the study of [26] considers limited fixed-order models (four-order polynomials) encoding limited fixed-size sets of parameters (ten LSF values). However, the study of [26] provides an inspiring basis for applying our LT modeling approach to very low bit-rate speech coding, as discussed at the end of this paper.

This paper is organized as follows. The proposed long-term model is described in Section II. The complete analysis-modeling-synthesis process is presented in Section III, including the description of the fitting algorithm. The perceptual criteria are presented in Section IV, for both amplitude and phase modeling. Experiments and results are given in Section V, evaluating the proposed approach in terms of modeling accuracy, synthesis signal quality, and data compression. Finally, the interest of such models for speech coding and watermarking is discussed in Section VI.

II. Long Term Models

As mentioned before, we suppose that the signal is already segmented into voiced and unvoiced parts by some usual voiced/unvoiced classifier (not described here). For each partial i, $1 \leq i \leq I$, we consider then the problem of separately modeling the time-trajectory of amplitude parameters $A_i(n)$ and phase parameters $\theta_i(n)$ of (1) over an entire voiced section of speech $s(n)$, running arbitrary from $n = 0$ to N. Different kinds of models can be proposed for this task. The problem of choosing appropriate models and comparing their performances is more extensively discussed in Section V. We only consider in this paper the discrete cosine model (DCM), which is a linear combination of cosine functions

$$\hat{A}_i(n) = \sum_{p=0}^{P_i} c_{ip} \cos\left(p\pi \frac{n}{N}\right) \tag{3}$$

where P_i is the order of the model, and the P_i+1 coefficients c_{ip} are all real. Such a model is known to be efficient in capturing the variations of a signal (e.g., when directly applied to the signal samples as for the DCT, or on its spectral envelope as in [28], [29]). Thus, it should be well suited to capture the global shape of sinusoidal parameter trajectories.

In the case of phase modeling, a linear term is added to the model, resulting in a linear + DCM (LDCM) model

$$\hat{\theta}_i(n) = \sum_{p=0}^{P_i} c_{ip} \cos\left(p\pi \frac{n}{N}\right) + c_{i(P_i+1)}n. \tag{4}$$

This linear term is quite useful for modeling the basic linear background shape of the phase trajectories, which results from the integration in time of the frequency trajectories. Thus, the cosine functions are used to model the variations of the phase trajectories around this basic linear shape. Of course, these variations are closely related to the frequency variations, as discussed in more detail in Section VI-B.

III. Analysis, LT Modeling, and Synthesis

A. Analysis

Although the analysis step aims at providing the set of amplitude and phase measures to be long-term modeled, it is still a short-term process as in typical analysis-modeling-synthesis

[1] This would be also true in the short-term approach, e.g., the partial tracking process of [7] is largely simplified in the harmonic case.

processes. The experiments described in this paper were conducted with a pitch-synchronous analysis. The voiced sections of the processed speech signals are first pitch-marked using Praat.[2] This means each period of signal is automatically time-labeled and used as an analysis frame.[3] The fundamental frequency ω_0^k at the center of the kth period is directly given by the inverse of the period length. Then, given the fundamental frequency, the amplitudes A_i^k and phases θ_i^k of the harmonics at the center of each period are estimated using the procedure of George and Smith in [19]. The estimation is based on an iterative minimum mean square error (MMSE) fitting of the harmonic model with the signal. This analysis method has been shown to provide very accurate parameter estimation with very low computational cost.

The phase estimation provides modulo 2π values θ_i^k that must be unwrapped to correctly reflect the "true" phase trajectory that we want to model: fluctuations around an increasing linear background shape, resulting from the summation of positive frequency values in (2). The unwrapping is done by cumulate addition of M times 2π to each measured phase value, with M given by [7]

$$M = e\left[\frac{1}{2\pi}\left(\theta_i^k - \theta_i^{k+1} + i\left(\frac{\omega_0^k + \omega_0^{k+1}}{2}\right)L_k\right)\right] \tag{5}$$

where $e[x]$ denotes the nearest integer from x and L_k is the number of samples between the centers of analysis frames k and $k+1$. Since we used a pitch-synchronous analysis, M is most of the time (but not always) equal to the rank of the analyzed partial. In the following, we always deal with unwrapped phases, while keeping the same notation θ_i^k as before.

At the end of the analysis process, each section of K consecutive periods of voiced speech is represented by I sets of K amplitude parameters $\mathbf{A}_i = [A_i^1 \quad A_i^2 \quad \ldots \quad A_i^K]^t$, and I sets of K unwrapped phase parameters $\boldsymbol{\theta}_i = [\theta_i^1 \quad \theta_i^2 \quad \ldots \quad \theta_i^K]^t$, since we have one set of each parameter for each partial trajectory (^{t} denotes the transposed vector/matrix).

B. LT Model Parameters Estimation

LT modeling consists in replacing each set amplitudes and phases, respectively, by a reduced set of DCM and LDCM coefficients, respectively. We first give in this subsection the general principle of the coefficients calculation, given that the order of the model is known: this is done by a weighted MMSE (WMMSE) minimization process. We then present an algorithm to automatically estimate the optimal model order and the optimal weights for each section of modeled speech.

Let us denote by $\mathbf{N} = [n_1 \quad n_2 \quad \ldots \quad n_K]^{\mathbf{t}}$ the vector containing the sample index of the signal period centers, and $\mathbf{M}_i$ the $K \times P_i + 1$ matrix that concatenates the DCM terms evaluated at the components of $\mathbf{N}$

$$\mathbf{M}_i = \begin{bmatrix} 1 & \cos\left(\pi\frac{n_1}{N}\right) & \cos\left(2\pi\frac{n_1}{N}\right) & \ldots & \cos\left(P_i\pi\frac{n_1}{N}\right) \\ 1 & \cos\left(\pi\frac{n_2}{N}\right) & \cos\left(2\pi\frac{n_2}{N}\right) & \ldots & \cos\left(P_i\pi\frac{n_2}{N}\right) \\ \ldots & \ldots & \ldots & & \\ 1 & \cos\left(\pi\frac{n_K}{N}\right) & \cos\left(2\pi\frac{n_K}{N}\right) & \ldots & \cos\left(P_i\pi\frac{n_K}{N}\right) \end{bmatrix}. \tag{6}$$

Let us also denote by $\mathbf{C}_i = [c_{i0} \quad c_{i1} \quad \ldots \quad c_{iP_i}]^t$ the vector of LT model coefficients. When we use the LDCM instead of the DCM, $\mathbf{N}$ is concatenated to $\mathbf{M}_i$ so that we have an additional column of linear terms, and $\mathbf{C_i}$ has one additional entry $C_{i(P_i+1)}$. Now, the WMMSE estimation of $\mathbf{C}_i$ is given by minimizing the weighted mean square error between $\mathbf{M}_i\mathbf{C}$ and the parameters vector $\mathbf{V}_i = \mathbf{A}_i$ or $\mathbf{V}_i = \boldsymbol{\theta}_i$, depending on which parameter trajectory is modeled, over all possible vectors $\mathbf{C}$

$$\mathbf{C}_i = \arg\min_{\mathbf{C}\in R^{P_i+1}} \left\lfloor (\mathbf{M}_i\mathbf{C} - \mathbf{V}_i)^t\, \mathbf{W}\, (\mathbf{M}_i\mathbf{C} - \mathbf{V}_i) \right\rfloor \tag{7}$$

where $\mathbf{W}$ is a $K \times K$ diagonal matrix of (positive) weights that are iteratively adjusted in the algorithm presented next. Those weights depend on the perceptual criteria presented in Section IV. It is important to note that, to improve modeling accuracy, they are defined along the time axis to give a relative importance to different time regions.

Since the modeling process intrinsically aims at providing a reduction in data dimension, we assume[4] that $P_i + 1 < K$ and the optimal WMMSE coefficient vector is classically given by

$$\mathbf{C}_i = \left(\mathbf{M}_i^t\mathbf{W}\mathbf{M}_i\right)^{-1}\mathbf{M}_i^t\mathbf{W}\mathbf{V}_i. \tag{8}$$

As mentioned in the introduction, the shape of parameter trajectories can vary widely, depending on the length of the considered voiced section, the phoneme sequence, the speaker, the prosody, or the rank of the partial. Thus, the appropriate order of the LT model for these trajectories can also vary widely. It is therefore crucial to find a method to automatically adjust the order for each section of modeled speech, for each kind of parameter, and for each partial. For this aim, we propose an iterative algorithm, which updates the order and the weights of the WMMSE process according to perceptual constraints. We give here the general form of the algorithm. The precise perceptual criteria for amplitude and phase modeling to be used in the algorithm are given in detail in Sections IV-A and B, respectively.

The general principle of the proposed algorithm is to give more weight in the WMMSE process of (8) to time regions where a given function $f(\mathbf{E}_i)$ [given later in (10) and (17)] of the modeling error $\mathbf{E}_i = \mathbf{M}_i\mathbf{C}_i - \mathbf{V}_i$ overcomes a given perceptually based threshold model $\mathbf{T}_i = [T_i^1 \quad T_i^2 \quad \ldots \quad T_i^K]^t$ (also specified later). Accordingly, smaller weights are given to time indexes where the error is under the threshold model (steps 4 and 5 of the algorithm below). Hence, for a given model order, the weighting vector is iteratively updated until the ratio R of

[2] [Online]. Available: http://www.praat.org

[3] Of course, standard techniques using a fixed-size sliding analysis window could also be used. In this study, the pitch-synchronous analysis has the advantage of providing a large amount of coherent measures (one per period of signal), thus facilitating the fitting of the LT model, as we will see in Section III-B.

[4] When the LDCT model is used for phase modeling, we have $C \in R^{P_i+2}$ in (7), and we must assume that $P_i + 2 < K$. Note that the proposed LT models can be efficiently exploited in very-low bit-rate speech coding if in practice P_i is generally significantly lower than K (see Sections V andVI).

the time regions where the error is below the masking threshold overcomes a user-defined target ratio $R_{\min}$. (close to 1, e.g., from 0.75 to 0.90, see Section V). If the condition $R > R_{\min}$ holds within a maximum number of iteration *Itermax*, the model order is decreased, otherwise it is increased (step 6). Then, the weighting process is iterated. *Itermax* is within a range of 10–20 (see Section V). The order P_i is initially set to the power of two closest to $K/4$, and the order update δP_i, initially set to $P_i/2$, is divided by two at each iteration. This allows a large increase of the speed of the algorithm. The algorithm stops when $\delta P_i = 0$ and the last P_i value for which $R \geq R_{\min}$ is retained.

Algorithm to be applied to either amplitude or phase measure sets.

1) For each time index $k \in [1, K]$ and each partial $i \in [1, I]$, calculate the associated threshold T_i^k (see Sections IV-A and B). Form the threshold trajectory $\mathbf{T}_i$. Then, for each partial i:
2) Initiate the order P_i to the power of two closest to $K/4$ and the order update δP_i to $P_i/2$.
3) Initiate a K-diagonal weight matrix $\mathbf{W}$ with all diagonal entries set to one. Then iterate from step 4 to step 6:
4) Calculate the LT model coefficients $\mathbf{C}_i$ with (8) and the associated modeling error $\mathbf{E}_i = \mathbf{M}_i\mathbf{C}_i - \mathbf{V}_i$.
5) Increase the weight vector $\mathbf{W}$ according to

$$\Delta\mathbf{W} = f(\mathbf{E}_i) - \mathbf{T}_i$$

$$\Delta\mathbf{W} \leftarrow \Delta\mathbf{W} + \min(\Delta W)$$

(so that $\Delta W$is always positive)

$$\mathbf{W} \leftarrow \mathbf{W} + \operatorname{diag}(\Delta\mathbf{W}/\max(\Delta\mathbf{W}))$$

6) Calculate the ratio R of negative elements in $f(\mathbf{E}_i) - \mathbf{T}_i$. If $R < R_{\min}$ and the maximum number of iterations *Itermax* is not reached, then go to step 4.
Else if $R \geq R_{\min}$, set $P_i \leftarrow P_i - \delta P_i$, set $\delta P_i \leftarrow \delta P_i/2$, and go to step 3.
Else if $R < R_{\min}$ and *Itermax* is reached, set $P_i \leftarrow P_i + \delta P_i$, set $\delta P_i \leftarrow \delta P_i/2$ and go to setp 3.

C. Synthesis

After calculation of the model coefficients, the synthesis is achieved by simply applying (3) for the amplitude trajectories and (4) for the phase trajectories, from $n = 0$ to N. Since the amplitudes must always be positive values, possible local negative values (which are generally very small) are set to zero. Then, (2) and (1) are applied to generate the synthesis signal. Note that since their frequency varies with time, the higher-rank partials can locally overcome the Nyquist frequency. In that case, amplitude values corresponding to this "no signal's land" are also set to zero during the analysis, modeling, and synthesis steps. In this paper, we experimented only the modeling of the ten first partials of voiced speech (see Section V), all of them always lying under the Nyquist frequency. Thus we do not deal with this problem. Also remember that, in this study, we only consider voiced parts of speech. The unvoiced sections were simply concatenated with the LT-modeled voiced sections with local overlap-add windowing to avoid audible artifacts [19].

IV. Perceptual Weighting for Model Fitting

We give in this section a presentation of the perceptual criteria used in the algorithm of Section III-B. We first present the criterion used for amplitude trajectories modeling and then the one used for phase trajectories modeling.

A. Perceptual Criterion for Amplitude Trajectories

In the case of amplitude trajectories modeling, we considered perceptual constraints based on the widely used frequency-domain amplitude masking model [30], [31]. Thus, we aim at estimating the order of the LT model so that it is the minimal value for which the modeling error power is almost always below the masking threshold, and thus is expected to be inaudible. This condition is a quite standard issue in perceptual speech coding (see [31] for a complete review), but the major point here is that the terms "almost always" evoke a constraint over time, and not only over frequency: in the present study, we model separately the amplitude trajectory of each partial, and the associated modeling error trajectory must lie under the trajectory of the masking threshold over time, on the entire considered section of K speech frames.

Therefore, for amplitude LT modeling, the first step of the proposed algorithm of Section III-B consists in calculating a model of the masking threshold time-trajectory over the modeled speech section, from the successive short-term values of this masking threshold. Since the modeled speech section is considered as quasi-harmonic, we used a simplified version of the masking threshold of the ISO standard [30] also depicted in [31]. We calculated the short-term masking threshold at each time index k, $k \in [1, K]$, directly from the amplitude spectrum A_i^k, $i \in [1, I]$. Thus, we only considered the additive contribution of tonal maskers and not the contribution of noise maskers [31]. Each individual tonal masker T_{ij}^k at frequency $\omega_j^k = j\omega_0^k$ has a masking contribution at frequency $\omega_i^k = i\omega_0^k$ given by

$$T_{ij}^k = P_j^k - 0.275B_j^k + S_{ij}^k - 6.025 \tag{9}$$

where P_j^k is the power of the masker at frequency ω_j^k, all powers being normalized in SPL [30], B_j^k is ω_j^k expressed in Barks [32], [33], and S_{ij}^k is a piecewise-linear function of P_j^k that assumes the well-known approximate triangular shape of the masking threshold model [31]. The global short-term masking threshold T_i^k at frequency ω_i^k is obtained from the linear-scale summation over j of all individual maskers T_{ij}^k within a 10-Bark neighborhood, also including the absolute hearing threshold contribution [31]. Then, to obtain the time-trajectory $\mathbf{T}_i$ of the threshold model for each partial i, the resulting values T_i^k are simply sorted along the time axis according to $\mathbf{T}_i = [T_i^1 \quad T_i^2 \quad \ldots \quad T_i^K]^t$. Note that this time model takes into account the influence of the other partials, since the short-term masking threshold model it is built on does.

Once the masking threshold time-model is available, it can be compared with the power of the modeling error over time, thus the function used in the algorithm of Section III-B is given by (square denotes the element-wise square function)

$$f(\mathbf{E}_i) = \frac{1}{2}\text{square}(\mathbf{A}_i - \hat{\mathbf{A}}_i) = \frac{1}{2}\text{square}(\mathbf{A}_i - \mathbf{M}_i\mathbf{C}_i). \tag{10}$$

B. *Perceptual Criterion for Phase Trajectories*

The problem of estimating the model order for phase trajectories with perceptual constraints is much more delicate than in the case of amplitude trajectories. Indeed, perceptual effects of phase distortion of sinusoidal signals are rather complex. Although they have been extensively studied (see a good recent review in [34]), they are more or less clearly identified only in the ideal stationary case. For example, Ploboth and Kleijn [34] or Kim [35] only considered the case of synthetic periodic vowels. At the same time, relatively poor attention has been paid to accurately encode the phase information in the modern speech/audio coding systems, compared to the magnitude information. This is because phase is generally considered of less perceptual importance compared to magnitude. However, as stressed by the authors of [34], "it is fair to state that no sinusoidal-modeling-based speech coders exist that provide transparent speech quality without the use of explicit information about the STFT phase spectrum." Thus, the recent study of Kim [36] aimed at exploiting the result of [35] in a perceptually driven bit allocation process for phase quantization, but was also limited to periodic vowels. It was also the case in [34], in the design of perceptually weighted phase vector quantizers, and the authors concluded that "the improvement [of their perceptual encoding of phase] compared to squared-error encoding is small and generally will not justify the additional computational effort required." On the other hand, other studies show the improvement of real speech coded quality when using perceptually based phase codebooks, see, e.g., [37].

Now, in all those studies, the signals are assumed to be stationary, at least on each short-term frame when real speech is considered, and phases are of the general form

$$\theta_i(n) = i\omega_0 \times (n - n_0) + \theta_i(n_0) \tag{11}$$

where n_0 is an arbitrary time index reference (possibly the pitch pulse event), and the fundamental frequency ω_0 is supposed to be constant on each frame. Therefore, perceptual studies and quantization techniques actually concern the *phase offset* $\theta_i(n_0)$, that is the constant term of phase that encodes the relative offset between the stationary harmonics. Now, it seems clear to us that, for real speech, phase is better described by (2) (with harmonic relationship between frequencies in the pseudoperiodic case) than by consecutive local equations like (11), since the frequencies are constantly evolving in time, more or less rapidly. That is why we claim that phase distortion due to the LT modeling (and possibly coding) process should rather be described as a frequency modulation of the original frequency trajectories, and perceptual effects should be analyzed from this point of view.

Therefore, we propose to introduce in our LT modeling of phase trajectory a perceptual criterion based on the control of the error between the frequency LT model and the original frequency trajectory. This frequency LT model is defined as the derivative of the phase LT model of (4)

$$\hat{\omega}_i(n) = -\frac{\pi}{N}\sum_{p=0}^{P_i} pc_{ip}\sin\left(p\pi\frac{n}{N}\right) + c_{i(P_i+1)}. \tag{12}$$

Then, we propose to estimate the order of the LT phase model as the lowest order that assumes that the absolute difference between the frequency LT model and the corresponding original (measured) frequency trajectory remains below a frequency modulation (FM) threshold trajectory.

For this aim, we chose to exploit and extend some results of [32] and [38] characterizing the perceptual thresholds for a single stationary sinusoid modulated by a sinusoidal modulation. In [32], [38], the FM threshold, i.e., the maximum deviation $\Delta\omega$ of the tone frequency for which the modulation remains inaudible, has been shown to be approximately proportional to the tone (carrier) frequency, for a given modulation frequency (if this latter is significantly lower than the former), with an almost constant lower bound. For example, for a modulation frequency of 4 Hz, and a carrier frequency ω greater than 500 Hz, the authors of [32] propose the model

$$\Delta\omega \approx \max(2\text{ Hz}, 0.0035\omega). \tag{13}$$

Now, the fundamental frequency of usual speakers can be seen as the carrier frequency of time-varying sinusoids, within the approximate range 100–300 Hz. This also stands for the different harmonics of the signal, with a proportional range. As the proposed LT model of phase (4) and its derivative (12) are intrinsically "smooth," the modeling error can be seen as a frequency modulation that depends on the model order but that is always small compared to the range of the fundamental. Therefore, we propose to adapt the stationary case model of (13) to the nonstationary case of LT modeling. For each partial i, and each frame k of the modeled speech section, the FM threshold model we propose to use in the LT approach is

$$\Delta\omega_i^k \approx \max\left(2\text{ Hz}, \alpha\omega_i^k\right). \tag{14}$$

In (14), α is a constant ratio for which the allowed range has been estimated through pilot perceptual listening tests. These tests have revealed that it could be tuned to a significantly higher value than in the stationary FM case of [32], [38] without major perceptual degradation. In the experiments of Section V-B, α is within the range of 2%–5%.

Finally, since we used the harmonic assumption and did not make specific measures of the frequencies for each partial in the analysis step of Section III-A, we replace ω_i^k in (14) by the multiple of the measured fundamental frequency ω_0^k. Thus the frequency masking threshold trajectory for partial i in the algorithm of Section III-B is finally

$$\mathbf{T}_i = \Delta\omega_i = [\,\Delta\omega_i^1 \quad \Delta\omega_i^2 \quad \ldots \quad \Delta\omega_i^K\,]^t \tag{15}$$

with

$$\Delta\omega_i^k \approx \max\left(2\text{ Hz}, \alpha i\omega_0^k\right), \qquad \text{for } 1 \leq k \leq K. \tag{16}$$

Accordingly, the modeling error trajectory in the same algorithm is here defined by

$$f(\mathbf{E}_i) = \text{abs}(\omega_i - \hat{\omega}_i) = \text{abs}(\omega_i - \mathbf{Q}_i\mathbf{C}_i) \tag{17}$$

where $\omega_i = [\,i\omega_0^1 \quad i\omega_0^2 \quad \ldots \quad i\omega_0^K\,]^t$, abs denotes the entry-wise modulus function, and $\mathbf{Q_i}$ is the matrix "derived" from $\mathbf{M_i}$, as shown in (18) at the bottom of the next page.

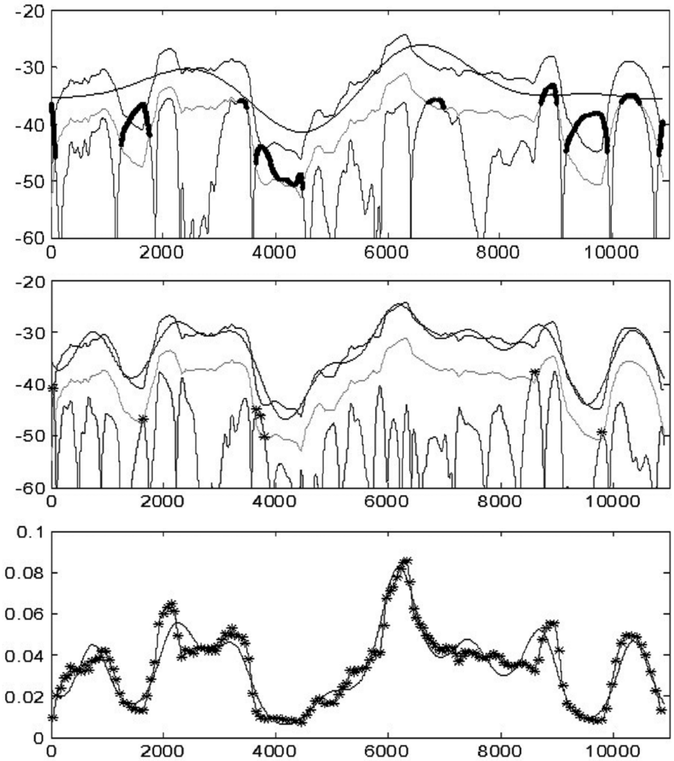

Fig. 1. Top and middle: Measured (upper raw curve) and DCM modeled (upper smooth curve) amplitude trajectory for the second harmonic of an all voiced male speech long sequence (11 000 samples at 8 kHz, $K = 157$); Masking threshold (light curve in the middle); Modeling error (lower curve); All curves are on log scale (decibels); $R_{\min} = 0.90$. Top: before convergence of Algorithm 1: $R < 0.70$, model order $P_i = 8$. Errors overcoming the masking threshold are marked by a star. Middle: after convergence of Algorithm 1: $R = 0.95$, order $P_i = 17$. Bottom: measured (smooth curve) and DCT modeled (stars) amplitude trajectory after convergence, on a linear scale.

V. Experiments

In this section, we describe the set of experiments that were conducted to test the presented LT modeling of amplitude and phase trajectories. We used 8-kHz continuous speech produced by 12 speakers (six male and six female speakers). About 3500 voiced segments of different sizes were modeled, representing more than 13 min of signal. For each section, the first ten harmonics were modeled. We first present in Section V-A a series of results concerning the amplitudes, and then we give in Section V-B a series of results on the phases.

A. LT Modeling of Amplitude Trajectories

1) LT Model Fitting: We first illustrate in Fig. 1 the behavior of the algorithm of Section III-B and the ability of the LT model to globally fit the amplitude trajectories. We plotted the trajectory of the second harmonic of a quite long sequence (≈ 1.4 s) of male voiced speech. We can see that the DCM model exhibits smooth trajectories around the amplitude measures. For this example, the final estimated DCM order is 17, while the number of measures is $K = 157$ ($R_{\min} = 0.90$). We found out that it is generally not necessary to force the modeling error to stay completely below the masking ratio (by setting $R_{\min} = 100\%$), for at least two reasons. First, lower ratios can provide high quality synthesis (e.g., $R_{\min} = 0.90$ to 0.75 according to the considered harmonic, see Section V-A2). Second, very local "strong modeling efforts" might result into a lower global fitting. Thus, it requires unnecessary computational efforts and model order increase to force small irregular portions of trajectory to lie under the masking threshold trajectory, especially when considering that such local irregularities can result from analysis errors.

Given a range of 0.75–0.90 for $R_{\min}$, it was generally found that less than ten iterations of the weighting process of the fitting algorithm were sufficient to test if the order model is appropriate or not (generally, after fast adaptive global shaping within a few iterations, there is a rather small evolution of the model from one iteration to the next). Also, it was found in practice that the algorithm generally converges toward an order value that is significantly lower than the number K of measures. This illustrates the ability of the proposed LT model to intrinsically allow data compression by efficient dimension reduction (see Section V-A3 for quantitative results).

2) Informal Listening Tests: To qualitatively assess the perceptual effects of the amplitude LT modeling, informal listening tests were conducted on signals of the test database. In this section, the signals were synthesized by applying the perceptually weighted LT model on amplitudes and by linearly interpolating the phase measures (as in [22]) before applying (2) and (1). This was because we wanted to separate the effects of amplitude LT modeling and phase LT modeling, the latter being tested in Section V-B. Also, remember that the LT model is applied only on the first ten harmonics. The amplitudes of the other harmonics are interpolated linearly. For comparison, we also synthesized reference signals with short-term linear interpolation of the measured amplitudes of all harmonics (and also linear interpolation of the phases).

Two subjects with normal hearing extensively listened to the synthesized signals using a high-quality PC soundcard and Sennheiser HD280 headphones. First, the perceptual difference between original and synthesis signals is quite low. Second, the main result of these tests is that *the long-term amplitude modeling generally provides a synthesis quality identical to the*

$$\mathbf{Q}_i = \begin{bmatrix} 0 & -\frac{\pi}{N}\sin\left(\pi\frac{n_1}{N}\right) & -\frac{2\pi}{N}\sin\left(2\pi\frac{n_1}{N}\right) & \dots & -\frac{P_i\pi}{N}\sin\left(P_i\pi\frac{n_1}{N}\right) & 1 \\ 0 & -\frac{\pi}{N}\sin\left(\pi\frac{n_2}{N}\right) & -\frac{2\pi}{N}\sin\left(2\pi\frac{n_2}{N}\right) & \dots & -\frac{P_i\pi}{N}\sin\left(P_i\pi\frac{n_2}{N}\right) & 1 \\ \dots & \dots & \dots & & & \dots \\ 0 & -\frac{\pi}{N}\sin\left(\pi\frac{n_K}{N}\right) & -\frac{2\pi}{N}\sin\left(2\pi\frac{n_K}{N}\right) & \dots & -\frac{P_i\pi}{N}\sin\left(P_i\pi\frac{n_K}{N}\right) & 1 \end{bmatrix} \quad (18)$$

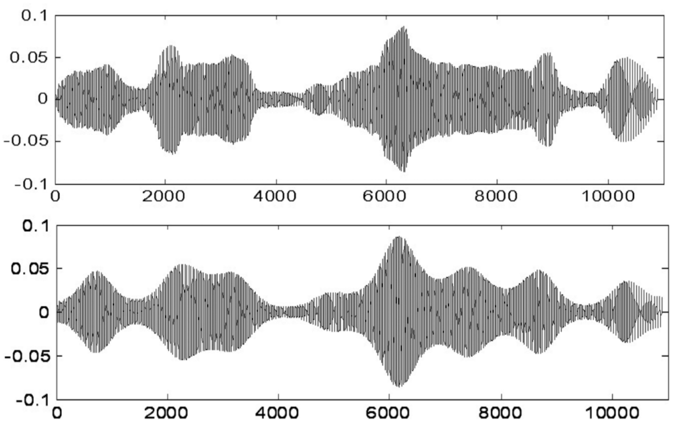

Fig. 2. Synthesized harmonic component corresponding to Fig. 1: amplitude short-term linear interpolation (top) and order-17 DCT model of Fig. 1 (bottom). Phase measures are linearly interpolated in both cases.

one obtained with the short-term linear interpolation of the measured amplitudes. In other words*, the signals synthesized with both ST and LT amplitude models cannot be distinguished.* Moreover, this result was possibly observed even for quite low model orders compared to the number of measures, e.g., for rates of about 20–30 coefficients/s depending on the considered harmonic and speaker gender. In fact, perceptual transparency between LT and ST modeling seems to be guaranteed as long as the modeling error power globally lies below the masking threshold, even if it locally overcomes this threshold: a value of $R_{\min} = 0.75$ seemed reasonable for most harmonics to ensure transparent quality compared to short-term modeled signals, although the two first harmonics seemed to possibly require a higher value (e.g., 0.90). This flexibility explains that the signal waveform shape may be significantly modified by the LT model without major perceptual effects (see Fig. 2). This demonstrates the efficiency of the LT version of the perceptual masking model and suggests that the corresponding LT amplitude model should be exploited in very low bit-rate speech coders.

3) Data Compression Gain: This subsection provides a first quantitative assessment of possible coefficient rates. For this, we selected the two values $R_{\min} = 0.75$ and $R_{\min} = 0.90$, after the listening tests of Section V-A1. We separated the results obtained for female and male voices, mainly because of their different ranges of fundamental frequency. Fig. 3 displays the model order value as a function of the length of the modeled speech section, for all sections of our database, for harmonics 2 and 5 and for $R_{\min} = 0.75$. These plots illustrate the diversity of situations, e.g., long sections modeled with quite low orders versus short sections modeled with higher orders. Then, these order values for all voiced segments were summed and divided by the total length of the segments, so that we obtain a mean coefficient-rate. This was done for each harmonic from rank 1 to 10 and for $R_{\min}$ set to 0.75 and 0.90.

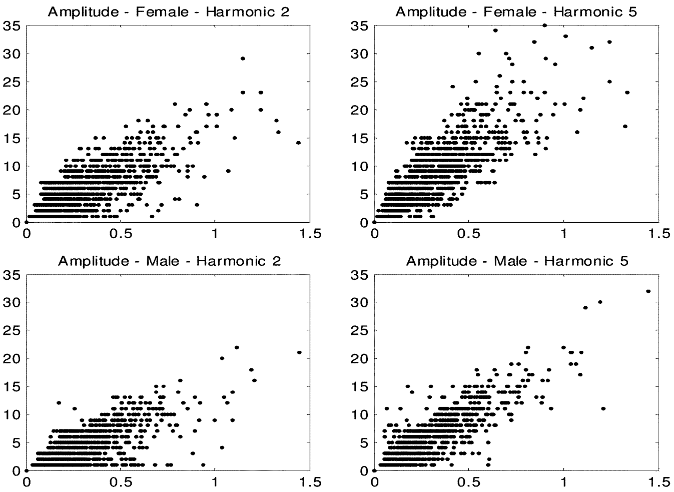

Fig. 3. LT model order value for amplitude modeling as a function of the length of the modeled speech section (in second). $R_{\min} = 75\%$. Top: female speakers. Bottom: male speakers. Left: harmonic 2. Right: harmonic 5.

TABLE I
RESULTS OF LT AMPLITUDE MODELING IN TERMS OF COEFFICIENT RATES (NUMBER OF COEFFICIENTS PER SECOND PER HARMONIC)

Harmonic	1	2	3	4	5	6	7	8	9	10
Fem. speech R_{min}=0.75	7.7	20.9	21.7	23.0	27.7	33.0	36.1	37.1	37.6	39.9
Male speech R_{min}=0.75	5.4	15.1	19.0	19.9	20.6	27.7	23.5	24.5	25.2	26.0
Fem. speech R_{min}=0.90	16.5	29.9	32.9	38.4	43.0	45.1	45.8	45.7	46.0	46.2
Male speech R_{min}=0.90	9.8	20.8	23.8	25.2	26.4	27.2	27.6	27.9	28.0	28.3

The results are gathered in Table I. They show that the mean order increases with the harmonic rank. This can also be seen on Fig. 3: the values for harmonic 2 are more spread out and shifted towards lower orders compared to the values for harmonic 5. This may happen because the amplitude trajectory generally becomes more complex as the rank of the harmonic increases, partly because of stronger noise/lower harmonic power in high-frequency regions, and corresponding difficulty to obtain accurate measures. This suggests again that constraints on order estimation (e.g., the $R_{\min}$ value) should be adapted to the harmonic rank. Unsurprisingly, the mean coefficient-rates are higher and more spread out for female speech than for male speech. This is coherent with the previous remark, since we must take into account the different pitch ranges of female and male speech. Thus, it is more costly to model a female speech harmonic than a male speech harmonic of the same rank, whereas female speech globally requires less resource since it contains less harmonics. This observation is coherent with general results on the gender dependency of sinusoidal speech coding efficiency. Eventually, note that the coefficient-rates are higher for $R_{\min} = 0.90$ than for $R_{\min} = 0.75$. This is obvious, since the order estimation directly depends on this parameter.

Now, by averaging the mean order values across the ten harmonics, taking $R_{\min} = 0.90$ for harmonics 1 and 2 and $R_{\min} = 0.75$ for harmonics 3 to 10 after the results of Section V-A2, we obtain global mean values of 30.3 and 21.7 coefficients/s per harmonic, for female and male speech, respectively. Therefore, if as much female speech as male speech is considered, the mean coefficient rate across gender and per harmonic would be 26. Comparatively, the mean number of short-term analysis frame per second (which is also the mean pitch value since the analysis was pitch-synchronous) was about 220 for female speech and 140 for male speech. Thus, *the LT model allows to divide*

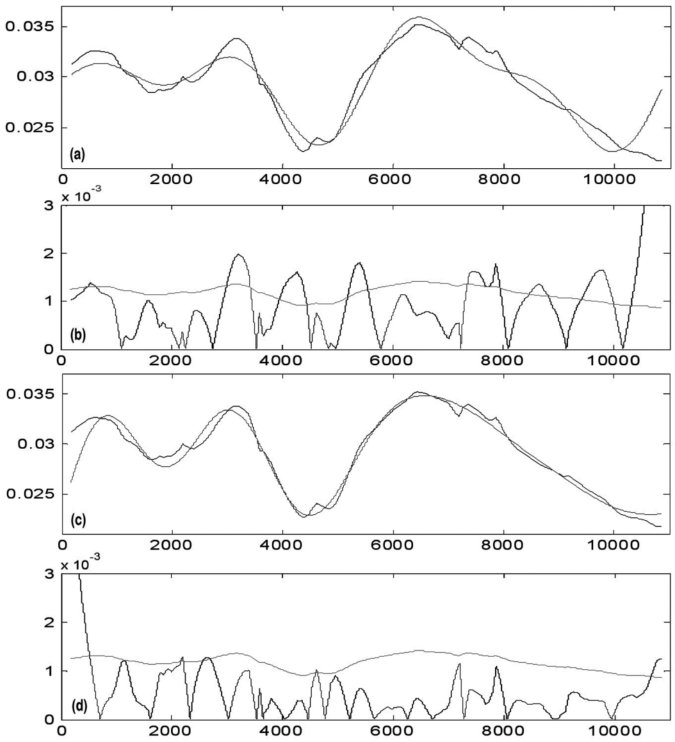

Fig. 4. (a) and (c) Measured (raw dark curve) and modeled (smooth light curve) frequency time-trajectory for the second harmonic of a long voiced male speech segment versus sample indexes (11 000 samples at 8 kHz, $K = 157$, same section as in Figs. 1 and 2). The frequency trajectory model, used in the perceptual matching algorithm, is the derivative of the phase trajectory model. (b) and (d) Frequency modulation threshold (upper light curve). Frequency modeling error (lower dark curve); $R_{\min} = 0.90; \alpha = 4\%$. (a), (b) Before convergence of the modeling algorithm: $R < 0.70$, model order $P_i = 9$. (c), (d) After convergence of the modeling algorithm: $R = 0.91$, model order $P_i = 11$.

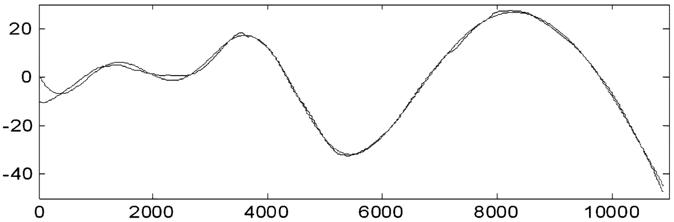

Fig. 5. Measured (raw dark curve) and modeled (smooth light curve) phase trajectory for the section of speech corresponding to Fig. 4(c) (male speech, 11 000 samples at 8 kHz, $K = 157, \alpha = 4\%, R = 0.91, P_i = 11$). The global linear term of the trajectories has been removed for better visualization.

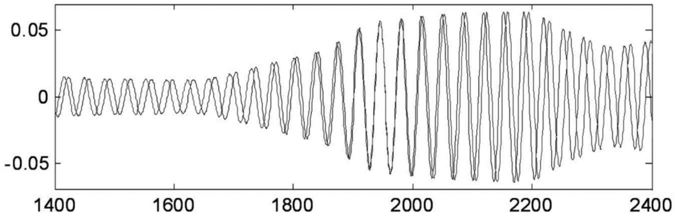

Fig. 6. Short-term and long-term phase synthesis signals corresponding to the 1400–2400 segment of Fig. 5. The amplitude measures are linearly interpolated. The LT synthesis signal is first in advance on the ST one, then it is late, as could be predicted from the phases of Fig. 5. Such dephasing can be interpreted as frequency modulation and is inaudible in this example.

the number of parameters by a factor of about 7 (at least for the ten first harmonics), compared to the short-term synthesizer using the pitch-synchronous measured amplitudes, while providing the same overall subjective quality (see Section V-A2). However, it may be fairer to compare the coefficient-rates with the ones of ST speech coders using usual fixed-size frame of about 10–20 ms, i.e., a range of 50–100 measures per second. In this case, *the coefficient rate is divided by a factor within the range* 2–4. Note that for such application as ultralow bit-rate speech coding, the model orders can be further decreased while preserving acceptable synthesis quality.

B. LT Modeling of Phase Trajectories

1) LT Model Fitting: As for the amplitudes, the LT model presents a rather good ability to globally fit the signal phase trajectories. Again, we arbitrarily selected the two values $R_{\min} =$ 0.75 and $R_{\min} =$ 0.90. In the case of phase modeling, we have the additional parameter α that controls the frequency modulation excursion (see Section IV-B). Figs. 4 and 5 represent the frequency and phase trajectories for the male voiced segment of Section V-A1. As before, the model exhibits smooth trajectories around the measures. For this example, the estimated order of the model is 11 (with $R_{\min} = 0.90$ and $\alpha = 4\%$), to be compared with the number of measures $K = 157$. Generally, the shape invariance of the synthesis signal is globally ensured for $R_{\min} = 0.90$, and α =2%–3%, since phase measures are well fitted by the model for the related orders. If the constraints are relaxed, local dephasing appears whereas the signal waveform is generally preserved (see Fig. 6, where $\alpha = 4\%$). This is because in this study *phase* measures and not frequency measures are modeled, even if the latter are used in the perceptive criterion.

2) Informal Listening Tests: As for amplitude modeling, informal listening tests were conducted on the signals of the test database. In this section, the signals were synthesized by applying the LT model on phase trajectories (of the first ten harmonics) and by linearly interpolating the amplitude measures. The two subjects mentioned in Section V-A2 listened to the newly synthesized signals. The main result of these new tests is that, as for amplitude LT modeling, *the LT phase model can provide a synthesis quality similar to the one obtained with short-term interpolation of the measured phases*. Because of the additional parameter α, it was difficult to extensively test all conditions. However, it can be reported that this result was obtained approximately for $R_{\min} \approx 0.80$ and $\alpha \leq 3\%$. For example, if α is set to 5%, a difference between LT and ST synthesized signals can be heard for some speech sections. The perceptual threshold value of α is thus difficult to evaluate, since it seems to depend on other modeling constraints (e.g., the value of $R_{\min}$), and possibly on additional signal characteristics (e.g., the amplitude of the modeled harmonic). However, it must be stressed that the range of perceptually acceptable values of α in this nonstationary LT modeling context is significantly higher than the values reported in [32] for stationary sinusoidal FM: about 0.03 in our case versus, e.g., 0.0035 in (13). This shows that the nonstationarity of real speech signals may provide an intrinsic masking effect on phase distortion, compared to the more

TABLE II
Coefficient -Rates of LT Phase Modeling (Number of Coefficients per Second per Harmonic), Female Speech, $R_{\min} = 0.75$

Harmonic	1	2	3	4	5	6	7	8	9	10
F. α= 2%	28.3	29.5	31.0	32.4	31.8	31.1	30.7	30.2	30.0	30.0
F. α= 3%	20.8	21.3	22.1	22.6	22.3	22.0	21.7	21.5	21.4	21.4
F. α= 4%	16.7	17.0	17.1	17.3	17.3	17.2	17.0	16.9	16.9	16.9
F. α= 5%	14.4	14.5	14.6	14.6	14.6	14.6	14.5	14.4	14.5	14.4

TABLE III
Coefficient -Rates of LT Phase Modeling (Number of Coefficients per Second per Harmonic) Male Speech, $R_{\min} = 0.75$

Harmonic	1	2	3	4	5	6	7	8	9	10
M. α= 2%	24.0	24.0	24.3	24.2	24.3	24.2	24.2	24.4	24.2	24.0
M. α= 3%	19.3	19.4	19.5	19.6	19.6	19.6	19.5	19.3	19.3	19.3
M. α= 4%	15.9	15.9	15.9	16.1	16.1	16.0	16.0	15.9	15.9	15.9
M. α= 5%	13.9	13.8	13.8	13.8	13.9	13.8	13.8	13.8	13.8	13.8

perceptually sensitive case of stationary signals. This seems to confirm previous similar results reported in [39]. Anyway, extensive formal listening tests should be conducted to quantify more precisely this point.

3) Data Compression Gain: In the case of phase modeling, the mean numbers of model coefficients per second over the complete test database and for the ten first harmonics are given in Table II for female speech and in Table III for male speech (both for $R_{\min} = 0.75$). As for amplitudes, the rates for female speech are greater than the rates for male speech, but the difference is much smaller than for amplitudes. For example, for $\alpha = 3\%$, the mean coefficient-rate across harmonics is 21.7 for female speech and 19.4 for male speech, resulting in a global rate across gender of 20.5 coefficients/s per harmonic. As could be expected, the rate increases when α decreases, but for the tested range of values, it always remains significantly lower than the number of measure values.[5]

Now, contrary to the amplitude results, for each α value, the coefficient-rates across harmonics remain very close. This is confirmed by the mean standard deviation of the order, calculated across harmonics for each section of modeled speech, and then averaged over the entire database. With $R_{\min} = 0.75$ and $\alpha = 3\%$, we found 0.6 for female speech and about 0.4 for male speech (low values are also found for other settings). This is due to the fact that, although the LT model fits *phase* trajectories, the perceptual criterion that drives the process is based on *frequency* trajectories fitting using the harmonic assumption. Therefore, we obtain for the different harmonics, a set of LT models that are almost in harmonic relationship (but not exactly, since it must jointly fit the phase measures).

Now, taking $R_{\min} = 0.75$ and $\alpha = 3\%$, we obtained above a mean rate of 20.5 coefficients/s per harmonic. Thus, *the LT model allow a gain-factor of approximately 7–10 on the number of phase parameters compared to the short-term-model-based synthesizer using the measured phases* (remember that the mean number of measured parameters is 220 for female speech and 140 for male speech). Compared to fixed-frame coders at 50–100 frames/s, *the gain is within the range* 2.5–5. Thus, as for amplitudes, the LT phase model could be useful for ultralow bit-rate speech coding. Also, in such application, the model orders could be further decreased while preserving acceptable synthesis quality.

VI. Discussion

A. Brief Summary of the Main Results

The proposed discrete cosine-based long-term model for sinusoidal speech amplitude and phase trajectories was shown to fit the measured parameter trajectories very well (at least for the ten first harmonics). An iterative algorithm provided efficient order estimation and perceptually relevant shaping of the model by using perceptual constraints on either amplitude or phase. As a result, the synthesized signals were of very good quality, at least as good as the signals synthesized with usual short-term models. Thus, the perceptually weighted LT model appears like a simple and efficient framework to study, mimic, and encode the time-dynamics of voiced speech.

B. Possibility for Other LT Models

Future work should include the implementation and test of other kinds of models. A polynomial model was also considered in our preliminary experiments on phase LT modeling [27]. It provided similar overall performances than the LDCM, while numerical instabilities due to the wide range of calculated values appeared when the order of the polynomial increased. This is why we did not consider this model in the present extended study. However, preliminary side tests let us think that this polynomial model could work well and even sometimes better than the DCM on short sections of voiced speech. Thus, future works will concern the use and comparison of other LT models, especially including the mixing of DCM and low order polynomial terms.

As mentioned in the introduction, the study of LT adequate models for unvoiced speech is also under consideration. We also plan to test a joint LT modeling of amplitude and phase, with combined perceptual criteria, since there may be some perceptual interference between them [37].

C. Application to Speech Coding

The proposed LT modeling was shown to provide synthesis signals of quality equivalent to short-term sinusoidal synthesis while providing efficient data compression through dimension reduction. Therefore, the presented method can be applied to very-low bit-rate speech coding, an application where the efficiency of the sinusoidal/harmonic model has already been shown [17]. LT models could lead to further significantly decrease the bit-rate of sinusoidal coders, although it would be at the expense of significantly increasing the encoding-decoding delay. This latter point is penalizing for telecommunication applications although it would not prevent the use of such coders in database storage and offline synthesis applications. Note that, since the presently proposed LT model is of variable-rate on variable-size frames, the resulting coders would also be of variable-rate and variable-delay. We are currently investigating in this direction, first addressing the problem of quantizing the LT model parameters or an equivalent representation of these

[5] For simplicity, extended results for $R_{\min} = 0.90$ are not presented, but it can be noted that the more drastic of our tested settings, $R_{\min} = 0.90$ and $\alpha = 2\%$ gives a mean rate across gender of 31.5 coefficient/s per harmonic.

parameters. Note that the quantization may benefit from the robustness to quantization of DCM coefficients as illustrated by the DCT transform.

Of course, the principle of such LT model-based coding should not be limited to sinusoidal context. It could be extended to other kinds of coders, e.g., LPC-based coders with LT modeling and coding of LSF parameters, as proposed in [26] with a fixed and limited multiframe approach. In fact, the LT approach could be extended to a wide range of model-based speech/audio coding techniques of different rates and quality. For example, for ultralow bit-rate speech coding, only the amplitudes and the fundamental frequency should be considered, and not the phases of the different harmonics. On the contrary, if one wants to accurately encode the phase values and not only the frequency values in order to preserve the signal waveform shape, only one additional parameter is needed from frequency trajectory modeling to phase trajectory modeling within the LT framework, since the frequency LT model is the derivative of the phase LT model.

D. Application to Speech Watermarking

Finally, we recently proposed an original speech watermarking process based on the sinusoidal model [39]. Watermarking consists in embedding additional data in a signal in an imperceptible way [40]. It is a technology of growing interest for copyrights and protection of data. In [39], we proposed to hide data within the dynamics of the frequency trajectories of the sinusoidal model of speech, by adequately modulating these trajectories. The watermarking process was shown to be efficient if the frequency trajectories that support the modulation were smooth enough, a property that may not be assumed by usual frame-by-frame interpolation schemes [21], [22]. The LT model presented in this paper is characterized by an intrinsic smoothness and should be used efficiently in the watermarking scheme. This point is also part of our future works.

REFERENCES

[1] J. A. Moorer, “Signal processing aspects of computer music—A survey,” *Comput. Music J.*, vol. 1, no. 1, pp. 4–37, 1977.

[2] ——, “The use of the phase vocoder in computer music applications,” *J. Audio Eng. Soc.*, vol. 26, no. 1/2, pp. 42–45, 1978.

[3] M. R. Portnoff, “Time-frequency representation of digital signals and systems based on short-time Fourier transform,” *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-28, no. 1, pp. 55–69, Feb. 1980.

[4] M. B. Dolson, “The phase vocoder: A tutorial,” *Comput. Music J.*, vol. 10, no. 4, pp. 14–27, 1986.

[5] J. B. Allen, “Short-term spectral analysis, synthesis, and modification by the discrete Fourier transform,” *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-25, no. 3, pp. 235–238, Jun. 1977.

[6] J.-C. Risset and M. V. Mathews, “Analysis of musical instrument tones,” *Phys. Today*, vol. 22, no. 2, pp. 23–30, 1969.

[7] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[8] X. Serra and J. O. Smith, “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic dec-omposition,” *Comput. Music J.*, vol. 14, no. 4, pp. 12–24, 1990.

[9] *“AudioSculpt User’s Manual,”* 2nd ed. IRCAM, Paris, France, 1996.

[10] K. Fitz and L. Haken, “Sinusoidal modeling and manipulation using lemur,” *Comput. Music J.*, vol. 20, no. 4, pp. 44–59, 1996.

[11] J. O. Smith and X. Serra, “PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation,” in *Proc. Int. Comput. Music Conf.*, San Francisco, CA, 1987, pp. 290–297.

[12] X. Serra, *Musical Signal Processing*. Lisse, The Netherlands: Swets & Zeitlinger, 1997, ch. Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122.

[13] S. Marchand and R. Strandh, “InSpect and respect: Spectral modeling, analysis and real-time synthesis software tools for researchers and composers,” in *Proc. Int. Comput. Music Conf. (ICMC’99)*, Beijing, China, 1999, pp. 341–344.

[14] A. Freed, X. Rodet, and P. Depalle, “Performance, synthesis and control of additive synthesis on a desktop computer using FFT-1,” in *Proc. Int. Computer Music Conf. (ICMC’93)*, Tokyo, Japan, 1993, pp. 98–101.

[15] J. W. Gordon and J. O. Smith, “A sine generation algorithm for VLSI applications,” in *Proc. Int. Comput. Music Conf. (ICMC’85)*, Vancouver, BC, Canada, 1985, pp. 165–168.

[16] J. O. Smith and P. R. Cook, “The second-order digital waveguide oscillator,” in *Proc. Int. Comput. Music Conf. (ICMC’92)*, San Jose, CA, 1992, pp. 150–153.

[17] R. J. McAulay and T. F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W. B. Kleijin and K. K. Paliwal, Eds. New York: Elsevier, 1995, ch. 4.

[18] T. F. Quatieri and R. J. McAulay, “Shape invariant time-scale and pitch modification of speech,” *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar. 1992.

[19] E. B. George and M. J. T. Smith, “Speech analysis/ synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 389–406, Sep. 1997.

[20] J. M. Grey, “An exploration of musical timbre,” Ph.D. dissertation, Dept. Music, Stanford Univ., Stanford, CA, 1975.

[21] Y. Ding and X. Qian, “Processing of musical tones using a combined quadratic polynomial phase sinusoid and residual signal model,” *J. Audio Eng. Soc.*, vol. 45, no. 7/8, pp. 571–585, 1997.

[22] L. Girin, S. Marchand, J. di Martino, A. Röbel, and G. Peeters, “Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals,” in *Proc. IEEE Workshop Applications Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, 2003, pp. 193–196.

[23] G. Richard and C. d’Alessandro, “Analysis/synthesis and modification of the speech aperiodic component,” *Speech Commun.*, vol. 19, pp. 221–244, 1996.

[24] D. W. Griffin and J. S. Lim, “Multiband-excitation vocoder,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 236–243, Apr. 1988.

[25] M. W. Macon and M. A. Clements, “Sinusoidal modeling and modification of unvoiced speech,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 557–560, Nov. 1997.

[26] S. Dusan, J. Flanagan, A. Karve, and M. Balaraman, “Speech coding using trajectory compression and multiple sensors,” in *Proc. Int. Conf. Speech Lang. Process.*, Jeju, Korea, 2004, CD-ROM.

[27] L. Girin, M. Firouzmand, and S. Marchand, “Long term modeling of phase trajectories within the speech sinusoidal model framework,” in *Proc. Int. Conf. on Speech & Language Proc.*, Jeju, Korea, 2004, CD-ROM.

[28] T. Galas and X. Rodet, “An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals,” in *Proc. Int. Comput. Music Conf. (ICMC)*, Glasgow, U.K., 1990, pp. 82–84.

[29] O. Cappé, J. Laroche, and E. Moulines, “Regularized estimation of cepstrum envelope from discrete frequency points,” in *Proc. IEEE Workshop Applications Signal Process. Audio Acoust. (WASPAA)*, 1995, pp. 213–216.

[30] *Information Technology–Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbits/s, Part 3: Audio*, ISO/IEC JTC1/SC29/WG11 MPEG, IS11172–3, 1992.

[31] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

[32] E. Zwicker and R. Feldtkeller, “Psychoacoustique: L’oreille récepteur d’information (French version of Das Ohr als Nachrichtenempfänger),” Masson, Paris, France, 1981.

[33] E. Zwicker and U. Zwicker, *Psychoacoustics Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.

[34] H. Pobloth and W. B. Kleijn, “Squared error as a measure of perceived phase distortion,” *J. Acoust. Soc. Amer.*, vol. 114, no. 2, pp. 1081–1094, 2003.

[35] D. S. Kim, “On the perceptually irrelevant phase information in sinusoidal representation of speech,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 900–905, Nov. 2001.

[36] ——, “Perceptual phase quantization of speech,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 355–364, Jul. 2003.

[37] O. Gottesman, “Dispersion phase vector quantization for enhancement of waveform interpolative coder,” in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Phoenix, AZ, 1999, pp. 269–272.

[38] L. Demany and C. Semal, “Detection thresholds for sinusoidal frequency modulation,” *J. Acoust. Soc. Amer.*, vol. 85, no. 3, pp. 1295–1301, 1989.

[39] L. Girin and L. S. Marchand, “Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials,” in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Montréal, QC, Canada, 2004, pp. 633–636.

[40] H. J. Kim, “Audio watermarking techniques,” in *Proc. Pacific Rim Workshop on Digital Steganography*, Kitakyushu, Japan, 2003.

Laurent Girin (M’06) was born in Moutiers, France, in 1969. He received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1994 and 1997, respectively.

In 1997, he joined the Ecole Nationale Supérieure d’Electronique et de Radioélectricité de Grenoble (ENSERG), Grenoble, France, where he is currently an Associate Professor in electrical engineering and signal processing. His research activity is carried out at the Institut de la Communication Parlée (Speech Communication Laboratory), Grenoble. It concerns speech processing, including speech analysis, modeling and synthesis, speech enhancement and audio/speech source separation, and a special interest in audiovisual speech processing.

Mohammad Firouzmand was born in Ghouchan, Iran, in 1965. He received the B.Sc. degree in electronic engineering from Gilan University, Rasht, Iran, in 1988, and the M.Sc. degree in biomedical engineering from Sharif University of Technology, Teheran, Iran, in 1992, and he is currently pursuing the Ph.D. degree at the Institut de la Communication Parlée (Speech Communication Laboratory), Grenoble, France.

From 1992 to 2003, he was Member of the Technical Staff and then a Project Leader at the Iran Communication Research Center, and then at the Biomedical Engineering Department of the Iranian Organization of Science and Technology (IROST).

Mr. Firouzmand was awarded by the Ministry of Science, Research, and Technology of Iran a scholarship to pursue the Ph.D. degree in Europe in 2003.

Sylvain Marchand was born in Pessac, France, in 1972. He received the M.Sc. degree in algorithmics and the Ph.D. degree in 2000, while carrying out research in computer music and sound modeling, from the University of Bordeaux 1, Talence, France, on 1996 and 2000, respectively.

He was appointed Associate Professor at the LaBRI (Computer Science Laboratory), University of Bordeaux 1, in 2001. He is particularly involved in spectral sound analysis, transformation, and synthesis.

Dr. Marchand is a member of Studio de Création et de Recherche en Informatique et Musique Electroacoustique (SCRIME), University of Bordeaux 1.