# Long Term Modeling of Phase Trajectories within the Speech Sinusoidal Model Framework

*Laurent Girin[(1)], Mohammad Firouzmand[(1)] & Sylvain Marchand[(2)]*

[(1)]ICP – INPG/Univ. Stendhal/CNRS
B.P. 25 - 38040 Grenoble France
{girin, firouz}@icp.inpg.fr

[(2)]SCRIME – LaBRI – Université Bordeaux 1
351, cours de la Libération – 33405 Talence France
sm@labri.fr

## Abstract

In this paper, the problem of modeling the trajectory of the phase of speech signal is addressed within the context of the sinusoidal model of speech. A global or long-term model of the trajectory of the phase of the partials is proposed for each entire voiced section of speech, contrary to standard models, which are defined on a frame-by-frame basis. The complete analysis-modeling-synthesis process is presented. We compare two basic long-term models, namely a polynomial and a DCT-based model, with classical (frame-by-frame) interpolation schemes, given that the analysis process is the same in all cases. Promising results are given and the interest of the presented models for speech coding and speech watermarking applications is discussed.

## 1. Introduction

Sinusoidal modeling of audio signals has been extensively studied since the eighties and successfully applied to a wide range of applications, such as coding or time- and frequency-stretching [1-5]. The signal is modeled as the sum of a small number $I$ of time-evolving sinusoids:

$$s(n)=\sum_{i=1}^{I} A_i(n)\cos\left(\theta_i(n)\right) \quad \text{with} \quad \theta_i(n)=\sum_{k=0}^{n}\omega_i(k)+\theta_i(0) \qquad (1)$$

The parameters of the model are the amplitudes $A_i(n)$, phases $\theta_i(n)$ and digital frequencies $\omega_i(n)$ (expressed in radians per sample) and are slowly evolving with time. An analysis-synthesis system based on such model usually requires the measurement of these parameters at the centers of consecutive signal frames, and then the interpolation of the consecutive measured values to reconstruct the entire signal. Amplitudes are generally interpolated linearly between frames, but for frequency and phase trajectories, the problem is slightly more complicated, since the frequencies are the time-derivatives of the phases, and the phase measures are defined modulo $2\pi$. If we want to reconstruct the signal with high fidelity regarding the waveform shape (the so-called shape-invariant property after [2]), not only measured frequencies but also measured phases must be respected. This implies to find models that match four constraints for each partial of each interpolated signal frame. One of the most famous models for phase interpolation that respects the shape-invariant property is the cubic polynomial interpolation proposed by McAulay and Quatieri [1]. Other models were proposed in [4-5].

In this paper, we propose a different approach to reconstruct the signal from the frequency and phase measures[1]. Instead of interpolating these values from one analysis frame center to the next, we propose to model the entire trajectory of each partial phase over each voiced section of speech with a single model. In other words, speech is first segmented into voiced and unvoiced parts, then the sinusoidal model is applied on each one of the voiced sections[2], and a single so-called "long term" (LT) model is used to represent the whole phase trajectory of a partial over the section. It is important to note that this voiced section can contain several phonemes (it can even be a complete sentence). That is why we also propose a method to automatically adjust the order of the model to the length of the section. In this paper, we propose and compare two possible LT models: a polynomial model and a linear+cosine model. They are described in section 2. The complete analysis-modeling-synthesis process is presented in section 3 and preliminary results are given in section 4. The interest of such models for speech coding and watermarking is discussed in section 5.

## 2. The long term phase models

As mentioned before, we suppose that the signal is previously segmented into voiced and unvoiced parts by usual voiced/unvoiced classifiers (not described here). We consider here the problem of modeling the trajectory of each partial phase over an entire voiced section of speech $s(n)$, running arbitrary from $n$=0 to $N$. We propose two different long-term models for the phase trajectories. The first model is a basic polynomial model:

$$\theta_i(n)=c_{i0}+c_{i1}n+c_{i2}n^2+\ldots+c_{iP_i}n^{P_i} \qquad (2)$$

while the second model is a combination of a linear term with a discrete cosine model, hence it is called linear+ discrete cosine model (LDCM):

$$\theta_i(n)=\sqrt{\frac{2}{N}}\sum_{p=0}^{P_i-1} c_{ip}w(p)\cos\left((n+\frac{1}{2})\frac{p\pi}{N}\right)+c_{iP_i}n \qquad (3)$$

[1] We do not deal with the amplitudes interpolation problem in this paper. Amplitudes are linearly interpolated as usual. LT amplitude models are currently being studied.

[2] The unvoiced sections are not considered in this paper. Other adequate models can be used for these sections, e.g. [6].

The factors $\sqrt{2/N}$ and $w(p)=1/\sqrt{2}$ if $p=0$ else $w(p)=1$ and the factor 1/2 inside the cosine are added to ensure perfect matching of the cosine part of the model with the standard discrete cosine transform (DCT). Note also that $P_i$ is the order of the polynomial model while it is the order minus 1 of the cosine model: this ensures to keep the same overall number of coefficients in the two cases.

The linear term is quite useful to model the basic linear background shape of the phase trajectories (which results from the integration in time of the frequency trajectories). Then, the cosine functions or the higher order polynomial terms are used to model the variations of the phase trajectories around this basic linear shape.

## 3. Analysis, modeling and synthesis

### 3.1. Analysis

The experiments described in this paper were conducted with a pitch-synchronous analysis. The signals were first pitch-marked by using the software Praat [7]. This means that the signals were considered quasi-harmonic and each period of signal was semi-automatically[3] time-labelled and used as an analysis frame. Thus, exploiting the pitch-marks, the fundamental frequency $\omega_0^k$ was directly given by the inverse of the period. Then, given the fundamental frequency, the amplitudes $A_i^k$ and phases $\theta_i^k$ of the harmonics at the center of each period were estimated by using the procedure used by George and Smith in [3]. The estimation is based on a classical minimum mean square error (MMSE) fitting of the harmonic model with the signal and it has been shown to provide very accurate parameter estimation with very low computational cost.

### 3.2. Phase unwrapping

The phase value estimation provides 2π–modulo values $\theta_i^k$ that must be unwrapped to correctly reflect the "true" phase trajectory that we want to model, that is an increase over time with fluctuations around a linear background shape, which results from the integration of the frequency values. The unwrapping is done by cumulate addition of $M$ times 2π to each measured phase value, with $M$ being the "unwrapping factor" of [1]:

$$M = e\left[\frac{1}{2\pi}\left(\theta_i^k - \theta_i^{k+1} + i\left(\frac{\omega_0^k + \omega_0^{k+1}}{2}\right)L_k\right)\right] \quad (4)$$

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.

### 3.3. Phase model parameters estimation

After the analysis process, each section of $K$ consecutive periods of voiced speech is represented by $I$ sets of $K$ amplitudes and unwrapped phases parameters (one set for each partial trajectory). Only the phase trajectories are modeled in this study, thus we consider the unwrapped phase measures sets, $i$=1 to $I$ ($^t$ denotes the transposed vector/matrix):

$$\boldsymbol{\theta}_i = \left[\theta_i^1 \theta_i^2 \ldots \theta_i^K\right]^t \quad (5)$$

Now we enter the very core of the presented study: we replace each set of phase parameters by a reduced set of either polynomial or LDCM coefficients. The fitting of the models with the measured unwrapped phase values is made by a standard MMSE minimization. Let us denote by $\boldsymbol{N} = [n_1\ n_2\ \ldots\ n_K]^t$ the vector of signal periods centers sample indexes, and $\boldsymbol{M_i}$ the matrix that concatenates the integer powers of the components of $\boldsymbol{N}$ when we use the polynomial model:

$$\boldsymbol{M}_i = \begin{bmatrix} 1\ n_1\ n_1^2 \ldots n_1^{P_i} \\ 1\ n_2\ n_2^2 \ldots n_2^{P_i} \\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ 1\ n_K\ n_K^2 \ldots n_K^{P_i} \end{bmatrix} \quad (7)$$

When we use the LDCM model, $\boldsymbol{M}_i$ is the matrix that concatenates $\boldsymbol{N}$ with the matrix of DCT terms evaluated at the components of $\boldsymbol{N}$:

$$\boldsymbol{M}_i = \sqrt{\frac{2}{N}} \begin{bmatrix} \frac{1}{\sqrt{2}}\cos\left(\left(n_1+\frac{1}{2}\right)\frac{\pi}{N}\right)\cos\left(\left(n_1+\frac{1}{2}\right)\frac{2\pi}{N}\right)\ldots\cos\left(\left(n_1+\frac{1}{2}\right)\frac{P_i\,\pi}{N}\right) & n_1 \\ \frac{1}{\sqrt{2}}\cos\left(\left(n_2+\frac{1}{2}\right)\frac{\pi}{N}\right)\cos\left(\left(n_2+\frac{1}{2}\right)\frac{2\pi}{N}\right)\ldots\cos\left(\left(n_2+\frac{1}{2}\right)\frac{P_i\,\pi}{N}\right) & n_2 \\ \ldots \quad \ldots \quad \ldots & \ldots \\ \frac{1}{\sqrt{2}}\cos\left(\left(n_K+\frac{1}{2}\right)\frac{\pi}{N}\right)\cos\left(\left(n_K+\frac{1}{2}\right)\frac{2\pi}{N}\right)\ldots\cos\left(\left(n_K+\frac{1}{2}\right)\frac{P_i\,\pi}{N}\right) & n_K \end{bmatrix}$$

Now the MMSE estimation of the coefficients vector $\boldsymbol{C_i} = [c_{i0}\ c_{i1} \ldots c_{iPi}]^t$ is found by minimizing the mean square error between $\boldsymbol{M_i C}$ and $\boldsymbol{\theta_i}$ over all possible vectors $\boldsymbol{C}$. Hence, it is given by:

$$\boldsymbol{C}_i = \left(\boldsymbol{M}_i^t \boldsymbol{M}_i\right)^{-1} \boldsymbol{M}_i^t \boldsymbol{\theta}_i \quad (11)$$

### 3.4. Model order tuning

Once the models and their coefficients estimation process are defined, it is necessary to find a method to automatically adjust the order of the models for each section of modeled speech. In this study, we attempted to tune the order so that a given mean signal-to-noise ratio (SNR) is achieved. This SNR is the ratio of the original signal power to the power of the difference between synthesis and original signals, both calculated over the considered modeled speech segment. Now, the model order is expected to generally depend on the length of the speech segment, as the longer the segment is, the more it can contain frequency variations. Thus, the basic idea is to first tune the model order for each segment of a training speech corpus so that a given minimum SNR is achieved[4]. This

[3] Semi-automatically refers to manual verification and minor local corrections after automatic extraction.

[4] In this preliminary study, for the purpose of simplicity, the model order is the same for each partial, while it could be differently adjusted for each partial in further studies.

can be done for several SNR ranges. Then, for each SNR range, a simple linear regression can be achieved relating the order to the segment length. After this, for each new segment of voiced speech to be modeled with a desired SNR range, the order of the model can be automatically (linearly) estimated from the segment length.

### 3.5. Synthesis

The synthesis is achieved by simply applying eq. 2 or eq. 3 depending on the chosen model, linearly interpolating the amplitudes between measured values and applying eq. 1. Remind that the whole analysis-synthesis process only concerns the voiced part of speech. In the following experiments, the unvoiced parts where kept as they are and concatenated with the modeled voiced parts with weighted overlap-add windowing to avoid audible artifacts [3].

## 4. Results

A set of experiments was conducted on speech signals consisting in 10-kHz sentences produced by 6 different speakers (3 males and 3 females). A total amount of 609 voiced segments of different sizes were used; representing nearly 2.5 minutes of voiced speech.

### 4.1. Model order estimation

In practice, we found out that for any speech segment to be modeled, the SNR generally grows with the model order, before stagnating near half the size of the measured parameter sets $K$, and decreasing approximately after $K$. This can be explained by an over-training phenomenon: when the degree of freedom of the models gets close to the number of constraints (that are minimizing the difference between $K$ modeled and measured values), these constraints are very well fitted, while for all other synthesis samples, the models do not capture well the signals characteristics. Thus, we cannot apply the idea of directly relating the model order to the SNR, since for short segments, we may have not enough phase measures to tune the models accurately. This will be tested in further studies, with analysis methods providing much more phase measures (e.g. sample-by-sample measures, as in the classical phase vocoder). In this study, we limited the maximal value of the order to $K/2$. In Fig. 1, we plotted the order of the polynomial model corresponding to the maximum SNR value obtained for each of the 609 voiced segments of the corpus, as a function of the segment length (similar graphs are obtained with the LDCT model). It is quite interesting to note that in most cases, the model order giving the maximum SNR is lower than $K/2$ (this is especially the case for long segments). The mean number of model coefficients per second per harmonic component over the complete corpus is 79 for the polynomial model and 91 for the LDCT model, providing respectively 17.5 and 17.9 dB mean SNRs, while the mean number of measured parameters is 200, providing 19.8 dB SNR when linearly interpolated using the standard short-term synthesis. Thus, *the LT models allow 120% gain on the number of phase parameters compared to the short-term coder using the measured phases, while only decreasing the SNR by around 2dB*. Quantization of the model parameters is a future trend of our work to apply the models to very low bit-rate speech coding. It is crucial to note that for such application, the model order (and thus the SNR) can be significantly decreased while preserving good subjective synthesis quality (see below).

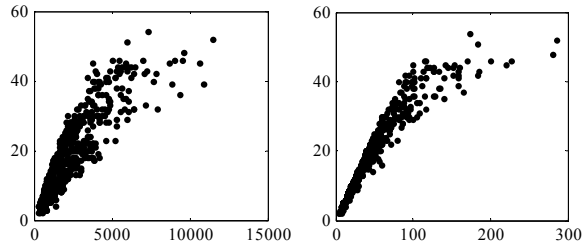


*Figure 1* – Polynomial phase model order vs. frame length in number of sample (left) or number of measures $K$ (right)

### 4.2. Original and modeled phase trajectories

To illustrate the ability of the models to follow the signal phase trajectories, we plotted in Fig.2 an example of such trajectories for a male voiced segment. We can see that the models exhibit smooth trajectories around the phase measures. Increasing the order allows to improve the fitting of the models with measures, and thus improve the SNR.

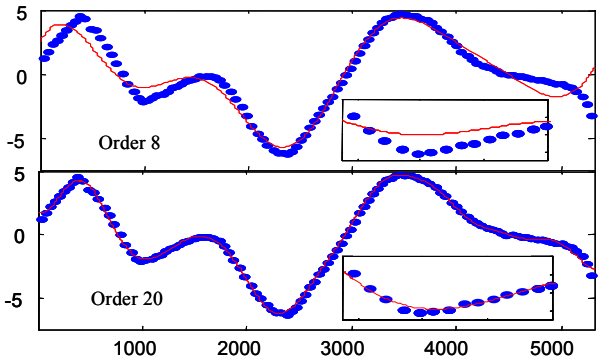


*Figure 2* – Measured (points – $K = 110$) and modeled (line) phase trajectory (polynomial model – order 8 and 20) for the first harmonic of a voiced male speech segment vs. sample indexes (sampling frequency = 10-kHz). The linear term of the trajectories has been removed for better visualization. The inserted rectangle is a zoom on samples 800 to 1400.

### 4.3. Synthesis signal shape

Globally the shape invariance of the synthesis signal is ensured for medium to high SNR values, since phase measures are well fitted by the models for the related orders. If the SNR (and so the model order) decreases, dephasing appears while the global signal waveform is often preserved. This is illustrated in Fig. 3.

### 4.4. Informal listening tests

Two subjects with normal hearing listened to the synthesized signals. First, the perceptual difference between original and synthesis signals is quite low, even if synthesized signals exhibits classical sinusoidal speech characteristics (e.g. the well-known “buzziness”). Second,

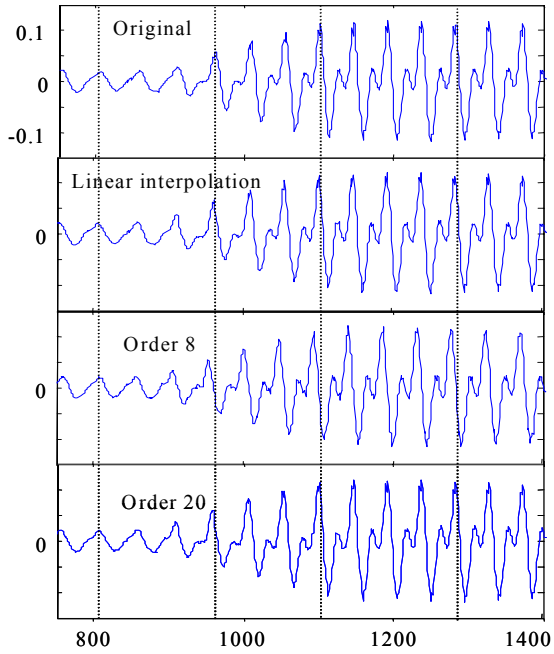


*Figure 3* – Original and synthesis signals corresponding to the 800-1400 segment of Fig. 2. The order 8 polynomial synthesis signal (SNR=2.3dB) is in advance on the others, as could be predicted from the modeled phase of Fig. 2. On the contrary, the order 20 synthesis signal (SNR=13.8dB) is synchronous with the original signal and the signal synthesized with (short-term) phase linear interpolation.

*the main result of these tests is that the models provide a synthesis quality similar to the one obtained with standard short-term (instead of long-term) interpolation of the measured phases*, as in [1, 2, 5]. It is very important to note that, even for quite low orders (e.g. 10 coefficients to model phase trajectories over several phonemes), the perceptual difference between long-term and short-term synthesis signals remains very low, although SNR drops significantly. This can be explained by the well-known relative unimportance of phase shifts in speech perception, a phenomenon that was recently confirmed in the sinusoidal model framework [8]. Again, this robustness of the models should be exploited in very low bit-rate high-delay speech coder.

## 5. Discussion

We proposed and tested two different long-term models for speech phase trajectories within the sinusoidal model framework: a polynomial and a linear+DCT model. Both were able to fit the local crucial phase variations around its global linear shape, even if the polynomial model appeared to be slightly more efficient than the LDCT model.

The presented approach can be applied to low bit-rate speech coding, an application where the efficiency of the sinusoidal model has already been shown [9]. The proposed models could lead to further decrease the sinusoidal coders bit-rate, although it would be at the cost of significantly increasing the encoding-decoding delay. We are currently investigating in this direction, addressing the problem of quantizing the LT model parameters together with LT modeling the amplitude trajectories.

Besides, we recently proposed an original speech watermarking process based on the sinusoidal model [10]. Watermarking consists in embedding additional data in a signal in an imperceptible way [11]. It is a technology of growing interest for copyrights and protection of data. In [10], 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 assured by usual frame-by-frame interpolation schemes [10][4]. The LT models presented in this paper are characterized by an intrinsic smoothness and should be used efficiently in the watermarking scheme. This point is also currently being investigated.

## 6. References


1. R. J. McAulay & T. F. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, **34**(4), 1986, pp. 744-754.
2. T. F. Quatieri & R. J. McAulay, Shape invariant time-scale and pitch modification of speech, *IEEE Trans. Signal Proc.*, **40**(3), 1992, pp. 497-510.
3. E. B. George & M. J. T. Smith, Speech analysis/ synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.,* **5**(5), 1997, pp. 389-406.
4. Y. Ding & X. Qian, Processing of musical tones using a combined quadratic polynomial phase sinusoid and residual signal model, *J. Audio Eng. Society*, **45**(7/8), 1997, pp. 571-585.
5. L. Girin, S. Marchand, J. di Martino, A. Röbel, & G. Peeters, Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals, *Proc. IEEE WASPAA*, New Paltz, 2003.
6. G. Richard & C. d'Alessandro, Analysis/synthesis and modification of the speech aperiodic component, *Speech Communication*, 19, 1996, pp. 221-244.
7. www.praat.org
8. D.S. Kim, On the perceptually irrelevant phase information in sinusoidal representation of speech, *IEEE Trans. Speech and Audio Proc.*, **9**(8), 2001, pp. 900-905.
9. R. J. McAulay & T. F. Quatieri, Sinusoidal coding, in *Speech coding and synthesis*, (W. B. Kleijn & K. K. Paliwal, eds), ch. 4, Elsevier, 1995.
10. L. Girin, L. & S. Marchand, Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials, *Proc. Int. Conf. on Acoustics, Speech & Signal Proc.,* Montréal, Canada, 2004
11. H. J. Kim, Audio watermarking techniques, *Proc. Pacific Rim Workshop on Digital Steganography*, Kitakyushu, Japan, 2003.