COMPARING THE ORDER OF A POLYNOMIAL PHASE MODEL FOR THE SYNTHESIS OF QUASI-HARMONIC AUDIO SIGNALS

Laurent Girin[1], *Sylvain Marchand*[2], *Joseph di Martino*[3], *Axel Röbel*[4], *and Geoffroy Peeters*[4]

[1] ICP – INPG/ENSERG
Université Stendhal
B.P. 25
F-38040 Grenoble, France
girin@icp.inpg.fr

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

[3] LORIA
Université H. Poincaré, Nancy 1
B.P. 239
F-54506 Vandœuvre, France
jdm@loria.fr

[4] IRCAM
1, place Igor-Stravinsky
F-75004 Paris, France
axel.roebel@ircam.fr
geoffroy.peeters@ircam.fr

ABSTRACT

Sinusoidal modeling has been successfully applied to a wide range of audio signal processing problems, such as coding or time and frequency stretching. While many methods have been proposed for the analysis part of the process, it seems that there is some general agreement concerning the synthesis part in the non-overlapping case: It is very often achieved by using the well-known McAulay-Quatieri method, which consists in an order 3 polynomial reconstruction of the phases of the sinusoidal model partials. In this paper, we compare this "classical" approach with both a simpler (order 1, that is linear interpolation) and a more complex (order 5) polynomial model for phase interpolation of quasi-harmonic signals. A gain has been measured in the signal-to-noise ratio at the synthesis stage, although the performance is limited by the amplitude model and by the imprecisions in the analysis stage.

1. INTRODUCTION

Sinusoidal modeling of audio signals has been extensively studied since the eighties and successfully applied to a wide range of audio signal processing problems, such as coding or time and frequency stretching [1, 2, 3, 4, 5]. The principle is to represent the signal as the sum of a small number P of sinusoids, given by:

$$s(n) = \sum_{p=1}^{P} A_p(n) \cos\left(\theta_p(n)\right) \tag{1}$$

$$\text{with} \quad \theta_p(n) = \theta_p(0) + \sum_{k=0}^{n} \omega_p(k) \tag{2}$$

The parameters of the model, respectively the amplitudes $A_p(n)$, phases $\theta_p(n)$, and (digital) frequencies $\omega_p(n)$ (expressed in radians per sample) are (slowly) evolving with time. An analysis-synthesis system based on such model requires the measurement of the parameters on adjacent / overlapping frames of signal and then the interpolation of the measured parameters to reconstruct the signal. While many methods have been proposed for the analysis part of the process (e.g. pick-peaking techniques on FFT spectrum in [1, 2] or minimum mean square error (MMSE) based analysis by synthesis in [3]), it seems that there is some "general agreement" concerning the synthesis part in the case where non-overlapping synthesis frames are used: It is generally achieved by using the well-known McAulay-Quatieri method [1], which consists in an order 3 polynomial reconstruction of the phases of the sinusoidal model partials[1]. The aim of this paper is to compare this "classical" approach with both a simpler (order 1, that is linear interpolation) and a more complex (order 5) polynomial model for phase interpolation. A quadratic phase model has been proposed by Ding and Qian in [6], while they have pointed out some drawbacks of the order 3 model. To our knowledge, the order 5 polynomial phase model has never been presented in the literature, while it presents *a priori* the advantage of better taking into account the evolution of the frequencies. This becomes possible because we are now able to measure the frequency derivative together with the phase and frequency during the analysis process [7, 8]. On the contrary, the order 1 (linear) model is simpler and would allow computational cost reduction in the systems. This paper reports the preliminary results that were obtained in the case of quasi-harmonic audio signals. This paper is organized as follows. In the next section, the three phase models are presented. Then the experiments are described in Section 3 and the results are given in Section 4 and discussed in the conclusion.

2. THE MODELS

The McAulay-Quatieri model for phase reconstruction of each signal partial[2] between the k-th and $(k+1)$-th synthesis frames consists of an order 3 polynomial, given by:

$$\theta(n) = \theta^k + \omega^k n + \alpha n^2 + \beta n^3 \tag{3}$$

where θ^k and ω^k respectively denote the phase and frequency of the partial measured at the junction of synthesis frames k and $k+1$ (which is chosen as the local origin $n = 0$). Assuming

[1] Amplitudes are generally linearly interpolated and we maintain this point in this paper.

[2] For simplicity sake, the partial subscript has been omitted in the following equations.

1. continuity of the phases and frequencies – which are the derivatives of the phases – at frame junctions,
2. unwrapping of the phase with a "maximally smooth" constraint on the phase model

leads to the model parameters α and β, given by:

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 3/N^2 & -1/N \\ -2/N^3 & 1/N^2 \end{bmatrix} \cdot \begin{bmatrix} \theta^{k+1} - \theta^k - \omega^k N + 2\pi M \\ \omega^{k+1} - \omega^k \end{bmatrix} \quad (4)$$

where N is the size of the synthesis frame, and M is the "phase unwrapping" integer factor given by:

$$M = e\left[\frac{1}{2\pi}\left((\theta^k - \theta^{k+1}) + (\omega^k + \omega^{k+1})\frac{N}{2}\right)\right] \quad (5)$$

where $e[x]$ denotes the nearest integer from x.

As mentioned above, this model ensures the continuity of the phases and frequencies at the frame junctions, but does not ensure the continuity of the frequency derivatives. Now, we suppose that we can also estimate the frequency derivatives at frame boundaries. In order to better take into account the frequency evolution of the partials, we propose to add the frequency derivatives continuity constraint, and to study the corresponding order 5 polynomial phase model, which is given by:

$$\theta(n) = \theta^k + \omega^k n + \frac{\psi^k}{2}n^2 + \alpha n^3 + \beta n^4 + \gamma n^5 \quad (6)$$

Note that, compared to the order 3 model, the first and second coefficients, respectively θ^k and ω^k do not change, since they still respectively represent the phase and phase derivative at $n = 0$. We denote by ψ the derivative of the frequency ω, that is the second derivative of the phase θ. Its value at $n = 0$ is ψ^k. The other constraints of the model on phase and frequency at frame junction are in this case:

$$\theta(N) = \theta^k + \omega^k N + \frac{\psi^k}{2}N^2 + \alpha N^3 + \beta N^4 + \gamma N^5 = \theta^{k+1} + 2\pi M \quad (7)$$

$$\omega(N) = \dot{\theta}(N) = \omega^k + \psi^k N + 3\alpha N^2 + 4\beta N^3 + 5\gamma N^4 = \omega^{k+1} \quad (8)$$

The additional constraint on the frequency derivative is given by:

$$\psi(N) = \ddot{\theta}(N) = \psi^k + 6\alpha N + 12\beta N^2 + 20\gamma N^3 = \psi^{k+1} \quad (9)$$

Solving the system of Equations from 7 to 9 leads to:

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 10/N^3 & -4/N^2 & 1/(2N) \\ -15/N^4 & 7/N^3 & -1/N^2 \\ 6/N^5 & -3/N^4 & 1/(2N^3) \end{bmatrix} \cdot \begin{bmatrix} \theta^{k+1} - \theta^k - \omega^k N - \frac{\psi^k}{2}N^2 + 2\pi M \\ \omega^{k+1} - \omega^k - \psi^k N \\ \psi^{k+1} - \psi^k \end{bmatrix} \quad (10)$$

The α, β, and γ coefficients can then be replaced in the expression of θ in Equation 6. For the order 5 model, M is then calculated by using the same criterion that the one used in [1], that is the maximal smoothing of the energy of the second derivative of the phase ($\psi = \ddot{\theta}$), by choosing the M that minimizes the function:

$$f : M \mapsto \int_0^N (\psi(n))^2 dn \quad (11)$$

In this case, we obtain:

$$M = e\left[\frac{1}{2\pi}\left((\theta^k - \theta^{k+1}) + (\omega^k + \omega^{k+1})\frac{N}{2} + (\psi^k - \psi^{k+1})\frac{N^2}{40}\right)\right] \quad (12)$$

Finally, we also implemented for comparison a linear phase model (order 1 polynomial) given by:

$$\theta(n) = \theta^k + \frac{\theta^{k+1} - \theta^k + 2\pi M}{N}n \quad (13)$$

with M given by Equation 5.

3. EXPERIMENTS

We have conducted series of tests for the three phase model orders on both synthetic and natural sound signals. The advantage of the synthetic signals is that, for each partial p, the phase ϕ_p, frequency ω_p, and frequency derivative ψ_p are known analytically, together with its amplitude A_p. These synthetic examples are tools for the investigation of the theoretical limits of the different order models in case of ideal analysis. For the natural sounds, these parameters were estimated using two analysis methods. The first one is pitch-synchronous (PS), whereas the second one is not.

3.1. Synthetic Examples

For all the synthetic examples, the sampling frequency is $F_s = 44100$ Hz (in the remainder, we denote by $T_s = 1/F_s$ the sampling period), the width of the synthesis frames is $N = 64$, and the total length of the sound in samples is $L = 1000N$. All these examples are harmonic sounds, made of $P = 20$ partials. The signals are also quantified using 16-bit precision, thus resulting in CD-quality sound examples.

We have chosen $\theta_p(0) = 0$, and we have $\psi_p(n) = \dot{\omega}_p(n)$. The expressions of the frequency ω_p and amplitude A_p have then to be defined for each example.

The first example is a perfectly stationary sound, where the fundamental frequency is $F_0 = 440$ Hz and:

$$A_p = 1/P \quad \text{and} \quad \omega_p = p2\pi T_s F_0 \quad (14)$$

The second example contains only linear variations. The amplitude is fading out while the fundamental frequency is raising (portamento from $F_0 = 440$ Hz to $2F_0$):

$$A_p(n) = (1 - n/L)/P \quad \text{and} \quad \omega_p = p2\pi T_s F_0 \cdot (1 + n/L) \quad (15)$$

It is clear that – unlike the previous examples – sinusoidal evolutions cannot be perfectly approximated by polynomials of finite degrees. The third example shows sinusoidal evolutions for the frequencies (vibrato), where the mean fundamental frequency is again $F_0 = 440$ Hz, and the vibrato depth and rate are respectively $F_1 = F_0/2$ and $F_v = 8$ Hz:

$$\omega_p(n) = p2\pi T_s \left(F_0 + F_1 \sin(2\pi F_v T_s n)\right) \quad (16)$$

The vibrato (sinusoidal variation of the fundamental frequency) was tested with and without tremolo (sinusoidal variation of the amplitude). Without tremolo, the expression of A_p is the same as in the constant case (see Equation 14). In the presence of tremolo, with a mean amplitude of $A_0 = 0.5$, and a tremolo depth and rate respectively set to $A_1 = A_0/2$ and $F_t = 8$ Hz, we have:

$$A_p(n) = \left(A_0 + A_1 \sin(2\pi F_t T_s n)\right)/P \quad (17)$$

The results obtained on these synthetic – but related to musical evolutions of the frequency and the amplitude, thus significant – examples can be found in Table 1. In this table, the signal-to-noise ratio (SNR) measures the energy ratio between the original signal s_o and the residual part (noise) obtained by subtracting the re-synthesis s_r from the original:

$$\mathrm{SNR}(s_o, s_r) = 10 \log_{10}\left(\sum_n s_o(n)^2 / \sum_n (s_o(n) - s_r(n))^2\right) \quad (18)$$

synthetic examples	1	3	5
constant	∞	∞	∞
linear	47.19	∞	∞
vibrato only	18.95	99.23	∞
vibrato+tremolo	19.20	76.21	76.23

Table 1: *SNRs in dB obtained on synthetic examples for the three phase models.*

The infinite symbol means that the two signals – original and re-synthesis – are identical (we reached the limit of the precision of 16-bit quantization for every sample). However, please note that the SNR is an imperfect perceptual metric. Since the SNR does not necessarily correlate with human perception, listening tests might have to be conducted in the near future. The case of constant parameters is perfectly handled by the three phase models, whereas linear evolutions (phase of order 2) require a polynomial order strictly greater than 1. The order 5 phase model better handles the sinusoidal evolutions of the frequency, and it is clear that the synthesis quality increases significantly with the phase model order in all cases of Table 1, except for the vibrato+tremolo case where the linear (order 1) interpolation we use for the amplitude within the synthesis frame is a bottleneck and should be enhanced (for example by an order 3 polynomial, provided that we can also estimate the derivative of the amplitude, see [7]).

3.2. Natural Sound Examples

We conducted experiments on a variety of quasi-periodic signals such as voiced speech and music (pieces of guitar and bass) originally sampled at 44100 Hz. The harmonic hypothesis was used, and for each frame k the frequencies ω_p^k are multiple of the fundamental ω_0^k – more precisely, we have $\omega_p^k = p\omega_0^k$ in Equation 2. Compared to the general case, this allows us to test the phase model without interfering with problems such as those encountered with the tuning of the partial tracker that is in charge of ensuring the continuity of the partial trajectories [1, 4]. In the harmonic case, harmonics of the same rank are simply connected to each other across the frames.

The experiments described in this paper were conducted in two analysis-synthesis conditions: pitch-synchronous (PS) or not.

3.2.1. Pitch Synchronous Analysis

In the PS condition, the signals were re-sampled at 10 kHz and pitch-marked previously to the analysis-synthesis process. This means that each period of signal was (semi-automatically[3]) time-labelled so that the analysis-synthesis process was conducted on successive periods of signal. The amplitudes and phases of the harmonics were estimated for each period of signal by using the procedure used by George and Smith in [3]. The estimation is based on a classical MMSE fitting of the harmonic model with the signal and it has been shown to provide very accurate parameter estimation. However, it requires to estimate first the fundamental frequency of the signal and the parameters estimation performance is quite dependent of the fundamental estimation. Exploiting the pitch-mark information, the fundamental was here directly given by the inverse of the period. For the order 5 polynomial phase model, the frequency derivatives were directly estimated in this preliminary study by taking the difference between two consecutive values of the frequency. Though quite coarse compared to much more sophisticated estimation algorithms (see below), this procedure was justified in the PS condition where the parameters related to frequency are estimated from each period of signal.

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

natural samples (PS)	1	3	5
speech (male)	21.15	23.55	23.87
speech (female)	25.75	27.52	27.83
singing voice	20.14	20.39	20.42
bass (short)	10.83	12.04	13.91
bass (long)	12.98	14.79	15.47
cello	18.00	19.24	19.67
electric guitar	10.83	12.04	14.39

Table 2: *SNRs in dB obtained for different signals and the three tested phase model polynomial orders (PS analysis method).*

3.2.2. Non Pitch Synchronous Parameter Estimation

Besides the pitch synchronous parameter estimation which is only suitable for monophonic sound signals, a more general analysis method estimating the partial parameters from the individual peaks has been used. To be able to evaluate the synthesis quality without having to solve the problem of partial tracking we still assume the sound sources to be harmonic and simply take the largest peak in a range of $\pm 0.4F_0$ around the theoretical partial frequency to estimate the partial parameters.

The frequency and frequency derivative (slope) of the partial related to individual spectral peaks is estimated relying on a recent reassignment technique [8]. Based on the frequency trajectory the optimal amplitude and phase can be derived by minimizing the error when the partial is subtracted from the signal. The fact that the frequency slope is considered for amplitude estimation significantly reduces the amplitude error for non stationary partials. For the current experiments we use a fixed window size of at least 5 periods of the minimum F_0 observed in the sound signal.

4. PRACTICAL RESULTS

Different speech and music signals[4] were processed through the analysis-modeling-synthesis process in the two (PS or not) conditions (see above) and with the three different polynomial orders for phase reconstruction. Speech signals consisted in continuous non-sense sentences with only voiced sounds (vowels and voiced/liquid consonants) uttered by a French male speaker and a female American English speaker. The two speakers were asked to produce sequences with a great range of fundamental variation. A piece of singing male voice was also tested with a quite more limited range of ω_0 values (3 notes). Musical signals consisted in pieces of electric and bass guitars. These pieces are solo performances by renowned "Metal" and Jazz musicians, and a large dynamic in ω_0 is guaranteed. Results are given in Table 2, in terms of signal-to-noise power ratios, where the noise is defined as the difference between the original and modeled signal.

The results of Table 2 indicate that the performances are increasing with the order of the polynomial phase model. The gains obtained by the proposed order 5 model over the classical order 3 model are approximately 0.3 dB for the speech signals, 0.7 dB for the bass and 2.3 dB for the guitar. The gain is very small for the singing voice (0.03 dB), reflecting the correlation between the increasing of the gain with the model order and the ω_0 dynamic of the signals. Indeed, the instrumental signals have the most important fluctuations in ω_0, then come the speech signals and then the singing – but quite stable – voice.

[4]URL: `http://dept-info.labri.fr/~sm/WASPAA03/`

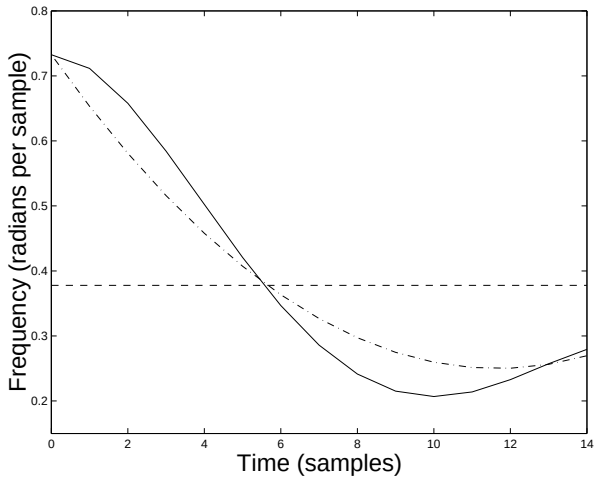

Figure 1: *Examples of frequency trajectories within one synthesis frame. The order 1 phase model leads to a constant frequency (dashed); the order 3 model makes the frequency appear as a parabolic segment (dash-dotted), while the order 5 model is more flexible (plain).*

Note that the SNR values for the instruments, which are quite low compared to the SNR values obtained with the speech signals, are due to the lower "voicing quality" of the signals. Some noise components (due to the friction the bow for the cello and the electric guitar distortion) may not be efficiently captured by the harmonic model. However, despite the extended range of SNR values across the different kinds of signal, it must be underlined that all synthesized signals were perceptually quite close to their corresponding originals.

Finally, it can be mentioned that the order 1 (linear) model logically provides the lowest performances (and except for the guitar piece, the SNR difference between order 1 and order 3 is always significantly larger than the SNR difference between order 3 and 5). However, it has the advantage of computational simplicity and may provide sufficient quality in the case of signals with small ω_0 dynamic as illustrated by the results on the singing voice signal.

Examples of the behavior of the frequency within a synthesis frame for the three phase models are illustrated in Figure 1.

The same experiments were then conducted with the non pitch synchronous analysis method (see Section 3). Very similar results were found, as shown in Table 3. Again, the bass guitar (short extract) was analyzed – using two different analysis window sizes – as well as the cello. Note that the SNR values for the non PS case are smaller due to the fact that a larger analysis window and frame offset have been used. Moreover, the non PS analysis restricts the model to the harmonic part of the signal whereas the PS analysis, due to its reduced frequency resolution, will model the noisy part of the sound using the harmonic model too, which may lead to artifacts during sound manipulation.

Of course, the results are sensitive to the analysis method used for the extraction of the model parameters. That is the reason why we repeated our comparison of the phase models on two different analysis methods.

The fact that the order 5 phase model does not significantly increase the SNR might be – similar to the case of the synthetic sound with amplitude modulation – due to the fact that the performance is limited by the linear amplitude interpolation. This hypothesis will be investigated in further experiments.

natural samples (not PS)	1	3	5
bass (short, small window)	8.39	9.25	9.44
bass (short, large window)	8.71	9.56	9.76
cello	16.35	16.92	17.02
violin	17.68	17.91	17.94

Table 3: *SNRs in dB obtained for different signals and the three tested phase model polynomial orders (non PS analysis method).*

5. CONCLUSION

In this paper, the problem of phase interpolation by polynomial modeling for audio signal synthesis was studied. It was shown that, compared to the classical order 3 model, an order 5 model can afford a performance gain that may be significant in the context of high-quality synthesis of quasi-harmonic signals with notable fundamental dynamic (a SNR improvement from 0.5 to 2 dB was reported for such signals). This gain is obtained at the price of an increased complexity in terms of calculation and also parameter estimation: frequency derivatives must be estimated since they provide the additional information on harmonic trajectories. Perspectives generally concern a more global study that would connect more in details the analysis process with the synthesis quality. As a special point, the importance of frequency derivatives estimation may be precisely studied. Also, the comparison of the model orders in the case of signal transformations such as pitch and/or time scaling should be conducted in the near future. An enhanced model for the amplitude parameter has to be investigated as well.

6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[2] T. F. Quatieri and R. J. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, 1992.

[3] E. B. George and M. J. T. Smith, "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis / Overlap-Add Sinusoidal Model," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 389–406, 1997.

[4] D. O'Brien and A. Monaghan, "Concatenative Synthesis Based on a Harmonic Model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 11–20, 2001.

[5] X. Serra, *Musical Signal Processing*, ser. Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997, ch. Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122.

[6] Y. Ding and X. Qian, "Processing of Musical Tones Using a Combined Quadratic Polynomial-Phase Sinusoid and Residual (QUASAR) Signal Model," *Journal of the Audio Engineering Society*, vol. 45, no. 7/8, pp. 571–585, 1997.

[7] M. Lagrange, S. Marchand, and J.-B. Rault, "Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model," in *Proceedings of the Digital Audio Effects (DAFx) Conference*, Hamburg, Germany, 2002, pp. 59–64.

[8] A. Röbel, "Estimating Partial Frequency and Frequency Slope Using Reassignment Operators," in *Proceedings of the International Computer Music Conference (ICMC)*, 2002, pp. 122–125.