PERCEPTUALLY WEIGHTED LONG TERM MODELING OF SINUSOIDAL SPEECH AMPLITUDE TRAJECTORIES

Mohammad FIROUZMAND & Laurent GIRIN

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

ABSTRACT

In this paper, the problem of modeling the trajectory of the amplitudes of speech signals is addressed within the context of the sinusoidal model of speech. A long-term model of the trajectory of the amplitude 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 a DCT-based long-term model with classical (frame-by-frame) interpolation schemes, given that the analysis process is identical in both cases. Perceptual constraints are taken into account since the distortion criterion in this approach is the level of modeling noise above the masking threshold. Promising results are given and the interest of the presented models for speech coding and 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 $A_i(n)$ and phases $\theta_i(n)$ and are slowly evolving with time (the digital frequencies $\omega_i(n)$, expressed in radians per sample, are the derivatives of the phases). 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. Phase measures are provided modulo 2π and must be unwrapped before interpolation [1]. Different models were proposed in the litterature for the frame-to-frame interpolation of phase parameters [1, 5, 6].

In a recent paper [7], we proposed a different approach to reconstruct the phase trajectories from its measures. Instead of interpolating these values from one analysis frame center to the next, we proposed to model the entire trajectory of each partial phase over each voiced section of speech with a single model. In other words, speech was first segmented into voiced and unvoiced parts, then the sinusoidal model was applied on each one of the voiced sections[1], and a single so-called "long term" (LT) model was used to represent the whole phase trajectory of a partial over the section. In this previous work, amplitudes were linearly interpolated as usual, in order to test separately the phase LT modeling. This latter was shown to provide a signal quality similar to the standard short term interpolation models, while concentrating the phase information in significantly fewer coefficients [7].

In this new paper, we deal with the dual problem of applying long term models to the amplitude trajectories of the partials along voiced sections of speech (while here classically interpolating the phase, e.g. with frame-by-frame linear interpolation). The LT model that we propose in this paper for amplitudes modeling is similar to one of the two models previously used for phase in [7]: it is a base of cosine functions. It is important to note that a modeled voiced section can contain several phonemes (it can even be a complete sentence). That is why, as for phase LT modeling, we propose a method to automatically adjust the order of the model for each amplitude trajectory. But the method is quite different in the case of amplitude modeling, since it is based on psychoacoustic constraints.

The paper is organised as follows. The LT model is 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 LT models for speech coding and watermarking is discussed in section 5.

2. THE LONG TERM AMPLITUDE MODEL

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 amplitude trajectory of each partial over an entire voiced section of speech $s(n)$, running arbitrary from $n = 0$ to N. We propose to use a discrete cosine model of the form:

$$A_i(n)=\sqrt{\frac{2}{N}}\sum_{p=0}^{P} c_{ip}\,w(p)\cos\left((n+\frac{1}{2})\frac{p\pi}{N}\right) \qquad (2)$$

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 model with the standard discrete cosine

[1] The unvoiced sections are not considered in this paper. Other adequate models can be used for these sections, e.g. [8].

transform (DCT). Such model (or transform) is known to be efficient to capture the slowly time-varying characteristics of a signal (e.g. on the speech signal samples themselves) and should be well suited to capture the global shape of sinusoidal parameter trajectories. In [7], a linear term was added to the DCT model to capture the linear background shape of phase trajectories. This is no more necessary for amplitudes. Thus, P_i denotes here the order of this model. Note that in [7] a polynomial model was also proposed for phase modeling, but was shown to be slightly less efficient than the linear+DCT model. Pilot tests have confirmed this result for amplitudes modeling, so that we concentrate this new study on the DCT model.

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 [9]. This means that the signals were considered quasi-harmonic and each period of signal was automatically time-labeled and used as an analysis frame. Thus, exploiting the pitch-marks, the fundamental frequency ω_0^k was directly given by the inverse of the period. Then, given the fundamental frequency, the amplitudes A_i^k and phases θ_i^k of the harmonics at the center of each period were estimated by using the procedure used by George and Smith in [4]. 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. Phase measures are provided modulo 2π and must be unwrapped by cumulate addition of M times 2π, M being the "unwrapping factor" of [1]. At the end of the analysis process, each section of K consecutive periods of voiced speech is represented by I sets of K amplitude and unwrapped phase parameters (one set for each partial trajectory, ^{t} denotes the transposed vector/matrix):

$$\boldsymbol{A}_i = [A_i^1 A_i^2 \ldots A_i^K]^t \text{ , } \boldsymbol{\theta}_i = [\theta_i^1 \theta_i^2 \ldots \theta_i^K]^t \qquad i=1 \text{ to } I \qquad (3)$$

3.2. Estimation of the LT amplitude model parameters

Now, amplitudes LT modeling consists in replacing each set $\boldsymbol{A}_i$ by a reduced set of DCT coefficients. The fitting of the model with the measured amplitudes is made by a standard MMSE minimization. Let us denote by $\boldsymbol{N} = [n_1\ n_2\ \ldots\ n_K]^t$ the vector of the sample indexes of the signal period centers, and $\boldsymbol{M}_i$ the matrix that concatenates the DCT terms evaluated at the components of $\boldsymbol{N}$:

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

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 $\boldsymbol{A}_i$ over all possible vectors $\boldsymbol{C}$. Hence, it is given by:

$$\boldsymbol{C}_i = \left(\boldsymbol{M}_i^t \boldsymbol{M}_i\right)^{-1} \boldsymbol{M}_i^t \boldsymbol{A}_i \qquad (5)$$

3.3. Synthesis

The synthesis is achieved by simply applying eq. 2 from n=0 to N, linearly interpolating the phases measures $\boldsymbol{\theta}_i$ and applying eq. 1. Note that since the partials frequency is varying with time, the higher rank partials can locally overcome the Nyquist frequency. In that case, amplitude values corresponding to this "no signal's land" can be set to zero during the analysis, modeling and synthesis steps. In this paper, we experimented only the modeling of the 4 first partials, all lying under the Nyquist frequency, thus actually we do not deal with this problem. Finally, 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 [4].

3.4. Model order tuning

Since the shape of the amplitude trajectories can vary widely, e.g. depending on the length of the voiced section, the phoneme sequence, or the rank of the partial, it is crucial to find a method to automatically adjust the order of the model for each section of modeled speech and for each partial. In this study, we considered perceptual constraints: the order is tuned so that the signal-to-noise ratio (SNR) is always over the signal-to-mask ratio (SMR). The SNR is defined as the ratio of the original partial power to the power of the difference between LT modeled and original partial. The SMR is defined as the ratio of the original partial power to the power of the perceptual frequency masking threshold [10]. In other word, the modeling error must always be under the masking threshold in order to be inaudible. This is a quite standard issue in speech coding [10, 11] but the major point in this new study is that the term "always" evokes here a constraint over time, and not only over frequency: we model separately the trajectory of each partial, and the associated modeling error trajectory must lie under the trajectory of the masking threshold over time. To achieve this goal, we propose to apply on each voiced section of K speech frames the following algorithm:

1) For each time index k and each set $\boldsymbol{A}_i^k$, $i \in [1,I]$ representing the speech spectrum magnitude at time k, calculate the associated global masking threshold $T^k(\omega)$ by using the model of [11] (also in [10] section II.F). Then, for each partial i:
2) Form the threshold trajectory $\boldsymbol{T}_i = \left\{T^k(\omega_i^k), k \in [1,K]\right\}$. Initiate the order P_i to *round*(K/4) and the order update dP_i to *round*(K/8), where *round* denotes the entire part.
3) Initiate a weight vector $\boldsymbol{W}$ of length K with all entries set to one. Then iterate the following process from step 4 to step 7:

4) Multiply element-by-element $\boldsymbol{W}$ with each column of $\boldsymbol{M}_i$ on the one hand and with $\boldsymbol{A}_i$ on the other hand, to obtain a weighted matrix $\boldsymbol{M}_i^W$ and a weighted vector $\boldsymbol{A}_i^W$.

5) Calculate the model by applying eq. 5 on the weighted data.

6) Increase the weights where the modeling error power $\boldsymbol{E}_i$ overcomes the masking threshold, according to (*square* and *max* respectively denotes the element-by-element square and maximum function):

$$\boldsymbol{E}_i = \frac{1}{2} square(\boldsymbol{A}_i - \boldsymbol{M}_i^W \boldsymbol{C}_i)$$

$$d\boldsymbol{W} = max(\boldsymbol{E}_i - \boldsymbol{T}_i, 0)$$

$$\boldsymbol{W} = \boldsymbol{W} + d\boldsymbol{W} / max(d\boldsymbol{W})$$

7) Calculate R the percentage of zero elements of $d\boldsymbol{W}$. If R is not over a given ratio R_{min} and the maximum number of iterations is not reached, then go to step 4.

   Else if $R \geq R_{min}$, decrease the model order $P_i = P_i - dP_i$, set $dP_i = round(dP_i/2)$ and go to step 3.

   Else if $R < R_{min}$ and the maximum number of iteration is reached, increase the model order $P_i = P_i + dP_i$, set $dP_i = round(dP_i/2)$ and go to step 3.

In the above algorithm, the maximum number of iteration and R_{min} are fixed arbitrary. Typically we can have respectively 20 and 90%. The algorithm stops either when $dP_i = 0$ or $P_i = K/2$ which is the limit of overtraining the models. Generally, the last value of P_i for which $R \geq R_{min}$ is retained. This ensures that the perceptual criterion is globally assumed. The dichotomic process for updating the model order allows to dramatically increase the speed of the algorithm. Finally, note that when the number of data is not sufficient to assume numeric stability (on short segments of voiced speech), both data and masking threshold can be linearly interpolated before modeling.

4. RESULTS

A set of experiments was conducted on speech signals consisting in 10kHz sentences produced by 6 different speakers (3 males and 3 females). A total amount of 610 voiced segments of different sizes were used; representing nearly 2.5 minutes of voiced speech.

4.1. Original and modeled amplitudes trajectories

Fig.1 illustrates the ability of the DCT model to globally fit the signal amplitude trajectories. We plotted the trajectory of the first harmonic of a long sequence (1.5 second) of female voiced speech. We can see that the model exhibits smooth trajectories around the amplitude measures. For this example, the order of the model is 26, to be compared with the number of amplitude measures $K = 408$ (see section 4.3.) As we can see from subplot a) and c), it is not necessary to force the modeling error to stay completely under the masking ratio (by fixing $R_{min} = 100\%$), since "very local strong modeling efforts" might result into a lower global fitting. In practice, as we will see in further sections, lower ratios can guarantee high quality synthesis (e.g. $R_{min} = 98\%$ to 75% according to the harmonic rank).

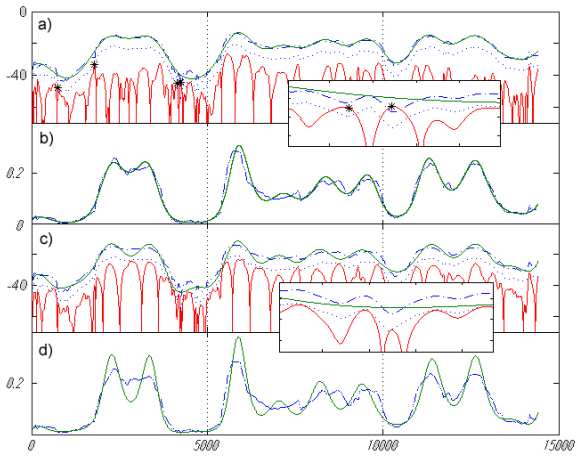

Figure 1 – Measured (dashed dotted) and DCT modeled (line – order 26) amplitude trajectory for the first harmonic of an all voiced female speech long sequence (15000 samples at 10kHz, $K = 408$) after a few iterations of the modeling algorithm; a)c) log scale; b)d) linear scale; The two lower curves on a) and c) are the masking threshold model (dotted) and the modeling error (line); a)b) $R_{min} = 98\%$ (errors overcoming the masking threshold are marked by a star); c)d) R_{min} is changed to 100% and a few more iterations are added; the inserted rectangle is a zoom on samples 4100 to 4400.

4.2. 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, thought synthesized signals exhibits classical sinusoidal speech characteristics (e.g. the well-known "buzziness"). Second, *the main result of these tests is that the long term model provides a synthesis quality identical to the one obtained with standard short-term linear or cubic interpolation of the measured amplitudes.* In other words*, the signals synthesized with both short or long term amplitude models cannot be distinguished.* Moreover, this result was observed even for quite low model orders compared to the number of measures (see e.g. in section 4.3.): it seems to be guaranteed as long as the modeling error (globally) lies below the masking threshold, despite the fact that the signal waveform shape may be significantly modified by the modeling process (see Fig. 2). This confirms the efficiency of the perceptual model and suggests that this perceptual robustness of the LT amplitude model should be exploited in very low bit-rate high-delay speech coders.

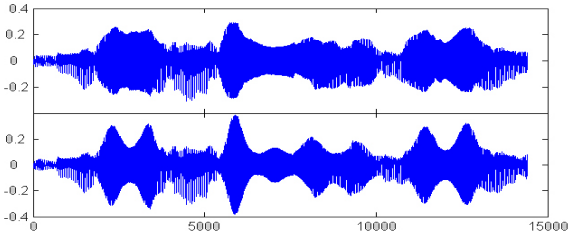

Figure 2 – Signal synthesized with amplitudes (short-term) linear interpolation and with the DCT model of Fig. 1 (only harmonic 1 is modeled, others are linearly interpolated).

4.3. Model order estimation and "compression gain"

To confirm this, we give in Table. 1 mean model order values obtained by summing the order values for the 610 voiced segments of our test corpus and dividing by the total length of the segments, and this for the 4 first harmonics. Perceptual informal tests led to chose R_{min}=98% for harmonics 1 and 2, and R_{min}=75% for harmonics 3 and 4, allowing to equilibrate the mean order between harmonics despite the fact that the amplitude trajectory generally becomes more complex as the harmonic rank increases. By averaging across the four harmonics, we obtain a mean value of 24.5 coefficients per second per harmonic, while the mean number of measured parameters is 201. Thus, *the LT model allows to divide the number of parameters by more than 8 (at least for the 4 first harmonics), compared to the short-term synthesizer using the measured amplitudes, while providing the same overall subjective quality*. Quantization of the model parameters is a future trend of our work to apply the method to very low bit-rate speech coding. It is crucial to note that for such application, the model order can be significantly decreased while preserving good subjective synthesis quality.

Harmonic Rate (number of coef./s)	1	2	3	4	Total
Weighted DCT model	20.8	33.3	20.7	23.5	98.3
Measures interpolation	201	201	201	201	804

Table 1 – Results in terms of coefficient rates

5. DISCUSSION

We proposed and tested a long-term model for speech amplitude trajectories within the sinusoidal model framework: a DCT perceptually weighted model. This model was shown to be able to fit the local amplitude variations of the lower rank harmonics of sinusoidal speech (from 1 to 4 in this study). Higher harmonics remains to be tested and part of our current work deals with this aim.

The presented approach, eventually associated with the phase LT modeling presented in [7], can be applied to very low bit-rate speech coding, an application where the efficiency of the sinusoidal model has extensively been shown [3]. The proposed models could lead to further decrease the sinusoidal coders bit-rate, though it would be at the cost of significantly increasing the encoding-decoding delay. Quantization of the model parameters and elaboration of a complete coder is a major part of our future work. Note that the quantization process may benefit from the well-known robustness of DCT coefficients already assessed in standard coding routines.

Besides, we recently proposed an original speech watermarking process based on the sinusoidal model [12]. Watermarking consists in embedding additional data in a signal in an imperceptible way [13]. It is a technology of growing interest for copyrights and protection of data. In [12], we proposed to hide data within the dynamics of the frequency trajectories of the sinusoidal model of speech, by adequately modulating these trajectories. The water-marking process was shown to be efficient if the trajectories that support the modulation were smooth enough, a property that may not be assured by usual frame-by-frame interpolation schemes [5][10]. The LT model presented in this paper is characterized by an intrinsic smoothness and should be used efficiently in the watermarking scheme, including when applied to the amplitude and not only the frequency trajectories. This point is also part of our future works.

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. R. J. McAulay & T. F. Quatieri, Sinusoidal coding, in *Speech coding and synthesis*, (W. B. Kleijn & K. K. Paliwal, eds), ch. 4, Elsevier, 1995.
4. 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.
5. 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.
6. 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.
7. L. Girin, M. Firouzmand & S. Marchand, Long term modeling of phase trajectories within the speech sinusoidal model framework, *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, South Korea, 2004
8. G. Richard & C. d'Alessandro, Analysis/synthesis and modification of the speech aperiodic component, *Speech Communication*, **19**, 1996, pp. 221-244.
9. www.praat.org
10. T. Painter & A. Spanias, Perceptual coding of digital audio, *Proc. IEEE*, **88**(4), pp.451-513.
11. ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3 Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s, Part 3: Audio, 1992.
12. L. Girin & 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
13. H. J. Kim, Audio watermarking techniques, *Proc. Pacific Rim Workshop on Digital Steganography*, Kitakyushu, Japan, 2003.