# Comparing Several Models for Perceptual Long-Term Modeling of Amplitude and Phase Trajectories of Sinusoidal Speech

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

[(1)]ICP (*Speech Comm. Lab.*) – INPG
46, av. Felix Viallet – 38031 Grenoble, France
{firouz, girin}@icp.inpg.fr

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

## Abstract

The so-called Long-Term (LT) modeling of sinusoidal parameters, proposed in previous papers, consists in modeling the entire time-trajectory of amplitude and phase parameters over large sections of voiced speech, differing from usual Short-Term models, which are defined on a frame-by-frame basis. In the present paper, we focus on a specific novel contribution to this general framework: the comparison of four different Long-Term models, namely a polynomial model, a model based on discrete cosine functions, and combinations of discrete cosine with sine functions or polynomials. Their performances are compared in terms of synthesis signal quality, data compression and modeling accuracy, and the interest of the presented study for speech coding is shown.

## 1. Introduction

Sinusoidal modeling of digital 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]. In such model, the speech signal is represented as the sum of a small number $I$ of time-evolving sinusoids (also called partials):

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

The amplitudes $A_i(n)$, phases $\theta_i(n)$ and digital frequencies $\omega_i(n)$ 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 measures at each sample time index to reconstruct the entire signal by applying (1) on the interpolated values. Amplitudes measures are generally interpolated linearly between frames and different models are proposed in the literature for the frame-to-frame interpolation of phase parameters [1][3][4].

In three recent papers [6][7][8], we proposed to model the entire trajectory of each partial phase [6][8] and amplitude [7][8] over each voiced section of speech with a single so-called Long-Term (LT) model. In other words, speech is first segmented into voiced and invoiced parts, then the sinusoidal parameters are measured along each voiced section on a short-term basis, and finally the LT model is used to represent the whole parameters trajectory over each entire voiced section.

In this paper, we present an important additional contribution to this approach: we focus on the specific problem of the choice of the LT model. Indeed, in [7][8] only one single model was considered: a sum of discrete cosine functions (plus a linear term in the case of phase modeling). In [6], we also dealt with a polynomial model (also considered in [9]) but a raw SNR-based criterion was used for phase model fitting. Since this criterion was replaced in [8] by a more pertinent perceptual criterion (see Section 3.3), the comparison of different LT models within the perceptual criterion framework has not been achieved yet, either for phase or amplitude trajectory modeling. This is the major point of the present study: we compare now the already assessed cosine-based model with three other models: a polynomial model, an hybrid cosine + sine model and an hybrid polynomial + cosine model. For this aim, we reuse the method and the perceptual criterions of [7][8] to automatically fit the models for each section of modeled speech.

This paper is organized as follows. The four LT models are presented in Section 2. The complete analysis-modeling-synthesis process is presented in Section 3. New experiments and results in Section 4 allow to compare the performances of the four competing models in terms of modeling accuracy, signal quality, and data compression. Finally, the interest of this study for speech coding is briefly discussed in Section 5.

## 2. The Long Term Models

As mentioned before, we suppose that the signal is previously segmented into voiced and unvoiced parts by usual V/UV classifiers (not described here). For each partial $i$, $1 \leq i \leq I$, we consider the problem of separately modeling the time-trajectory of the phase and amplitude parameters of (1) over an entire voiced section of speech $s(n)$, running from $n = 0$ to $N$. The first of the four proposed LT models is a polynomial model (PM):

$$\hat{\theta}_i(n)=c_{i0}+c_{i1}n+c_{i2}n^2+...+c_{iP_i}n^{P_i} \qquad (2)$$

The second model is a sum of cosine functions and a linear term, hence it is called linear + discrete cosine model (LDCM):

$$\hat{\theta}_i(n)=\sum_{p=0}^{P_i} c_{ip}\cos\left(\frac{p\pi n}{N}\right)+c_{i(P_i+1)}n \qquad (3)$$

The third model is a combination of linear term, discrete cosine and sine functions (LDCSM) ($P_i$ is assumed to be even here):

$$\hat{\theta}_i(n)=c_{i0}+\sum_{p=1}^{P_i/2}\left(c_{ip}\cos\left(\frac{p\pi n}{N}\right)+c_{i(p+P_i/2)}\sin\left(\frac{p\pi n}{N}\right)\right)+c_{i(P_i+1)}n \qquad (4)$$

The fourth model is a combination of polynomial and cosine functions (PDCM), in equal numbers ($P_i$ is assumed to be even):

$$\hat{\theta}_i(n)=c_{i0}+\sum_{p=1}^{P_i/2}\left(c_{ip}n^p+c_{i(p+P_i/2)}\cos\left(\frac{(p+P_i/2)\pi n}{N}\right)\right) \tag{5}$$

In each case, $P_i$ is a positive integer defining the total order of the model. The coefficients $c_{ip}$ are all real. The linear term is quite useful to model the linear background shape of the phase trajectories, which results from the time-integration of the frequency tracks. The higher-rank polynomials or trigonometric functions are useful to model the variations of the phase around this basic linear shape. Adding corresponding sine functions to the cosine amounts to allow these cosine for a free (model) phase offset. Thus, we want to test if allowing a phase offset is efficient compared to adding higher-rank cosine functions. And with the PDCM, we want to see if the sets of cosine and polynomial functions complete themselves well.

For the amplitudes, the models are the same, except that there is no linear term in (3)(4) since the amplitude trajectories do not systematically increase over time. Hence, we have for amplitude to compare the PM and PDCM with a discrete-cosine model (DCM) and a discrete cosine+sine model (DCSM).

## 3. Analysis, LT Modeling, and Synthesis

### 3.1. Analysis

The experiments described in this paper were conducted with the pitch-synchronous analysis previously used in [7][8] (see those papers for more details). Each period of signal was time-marked and used as an analysis frame. Thus, the fundamental frequency $\omega_0^k$ was given by the inverse of the period. The amplitudes $A_i^k$ and phases $\theta_i^k$ at the center $n_k$ of each period were estimated by the least mean square error (LMSE) procedure of [3]. This technique has been shown to provide accurate parameter estimation at very low computational cost. The phase values are provided 2π–modulo and are unwrapped by using the procedure of [1], to correctly reflect the "true" phase trajectories. Finally, for each section of $K$ speech periods, the analysis provides $I$ sets of $K$ amplitudes $\boldsymbol{A}_i=[A_i^1 A_i^2 \ldots A_i^K]^t$ and unwrapped phases $\boldsymbol{\theta}_i=[\theta_i^1\ \theta_i^2 \ldots\ \theta_i^K]^t$ at time indexes $\boldsymbol{N}=[n_1\ n_2\ \ldots\ n_K]^t$.

### 3.2. LT Model Coefficients Estimation

LT modeling consists in replacing each set of parameters by a reduced set of LT model coefficients by using LMSE regression. Let us denote by $\boldsymbol{M}_i$ the "LT model matrix". In the PM and DCM cases, $\boldsymbol{M}_i$ is the $K\times(P_i+1)$ matrix of general entry $m_{kp}=n_k^p$ and $m_{kp}=\cos(p\pi n_k/N)$, respectively. In the DCSM and PDCM cases, $\boldsymbol{M}_i$ results from the concatenation of general entries $m_{kp}=\cos(p\pi n_k/N)$ and $m_{kp}=\sin(p\pi n_k/N)$, and $m_{kp}=n_k^p$ and $m_{kp}=\cos(p\pi n_k/N)$, respectively. For the LDCM or LDCSM, $\boldsymbol{N}$ is concatenated to $\boldsymbol{M}_i$. The coefficients vector $\boldsymbol{C}_i=[c_{i0},c_{i1},\ldots,c_{iPi}]^t$ (with one more coefficient for the LDC(S)M) is found by minimizing the mean square error between $\boldsymbol{M}_i\boldsymbol{C}$ and $\boldsymbol{V}_i=\boldsymbol{\theta}_i$ or $\boldsymbol{V}_i=\boldsymbol{A}_i$ over all vectors $\boldsymbol{C}$. Hence, $\boldsymbol{C}_i$ is given by:

$$\boldsymbol{C}_i=\left(\boldsymbol{M}_i^t\boldsymbol{M}_i\right)^{-1}\boldsymbol{M}_i^t\boldsymbol{V}_i \tag{6}$$

The shape of amplitude and phase trajectories can vary widely, e.g. depending on the length of the section, the phoneme sequence, the speaker, the prosody, or the rank of the partial. Therefore, we proposed and tested in [7][8] a method to automatically fit the LT model and estimate its appropriate order for each section of modeled speech and for each partial. This method considers perceptual constraints for both amplitudes and phases LT modeling. The algorithm (to be applied to each voiced section of $K$ frames) is given below, the error $\boldsymbol{E}_i$ and perceptual threshold model $\boldsymbol{T}_i$ being defined below:

1. Initiate an arbitrary order $P_i$, e.g. the integer closest to $K/4$, and an order update $\delta P_i$, e.g. the integer closest to $P_i/2$. Initiate $R_{min}$ an arbitrary target ratio, typically 0.75–0.9
2. Initiate a weight vector $\boldsymbol{W}$ of length $K$ with all entries set to one. Then iterate from step 3 to step 6:
3. Multiply element-by-element $\boldsymbol{W}$ with each column of $\boldsymbol{M}_i$ on the one hand and with $\boldsymbol{V}_i$ on the other hand, to obtain a weighted matrix $\boldsymbol{M}_i^W$ and a weighted vector $\boldsymbol{V}_i^W$.
4. Calculate the LT model coefficients $\boldsymbol{C}_i$ by applying (5) to the weighted data.
5. Calculate the trajectory of the modeling error $\boldsymbol{E}_i$, the associated perceptual threshold model $\boldsymbol{T}_i$, and the difference $\Delta\boldsymbol{W}=\boldsymbol{E}_i-\boldsymbol{T}_i$. Increase the weight vector $\boldsymbol{W}$ according to:
   $\Delta\boldsymbol{W}\leftarrow\Delta\boldsymbol{W}+min(\Delta\boldsymbol{W})$ (so that $\Delta\boldsymbol{W}$ is always positive)
   $\boldsymbol{W}\leftarrow\boldsymbol{W}+\Delta\boldsymbol{W}/max(\Delta\boldsymbol{W})$
6. Calculate the percentage $R$ of negative elements in $\Delta\boldsymbol{W}$ (before adding $min(\Delta\boldsymbol{W})$). If $R<R_{min}$ and some maximum number of iterations is not reached, then go to step 3.

   Else if $R\geq R_{min}$, decrease the model order $P_i\leftarrow P_i-\delta P_i$, set $\delta P_i\leftarrow\delta P_i/2$ and go to step 2.

   Else if $R<R_{min}$ and the maximum number of iteration is reached, increase the model order $P_i\leftarrow P_i+\delta P_i$, set $\delta P_i\leftarrow\ \delta P_i/2$ and go to step 2.

The perceptual criterion for amplitude LT modeling is an adaptation of the frequency-domain masking threshold model of [10]: the values of the masking threshold model $T_i^k$ at each time index $n_k$ are resorted along the time axis, so that the time-trajectory of the threshold model $\boldsymbol{T}_i=[T_i^1\ T_i^2\ldots\ T_i^K]^t$ is obtained for each partial $i$ [7]. Here, $\boldsymbol{E}_i$ is actually the trajectory of the *power* of the modeling error $\boldsymbol{A}_i-\boldsymbol{M}_i\boldsymbol{C}_i$ along the time axis, and indeed, it must remain under the trajectory of $\boldsymbol{T}_i$.

For the phases, we proposed, discussed and tested in [8] a perceptual criterion based on a frequency modulation (FM) threshold model: for each partial, the absolute difference between the *derivative* of the LT model of phase and the corresponding frequency trajectory must stay under an FM threshold model. This latter is an adaptation of the (static) threshold identified in experiments on the detection of sinusoidal (fixed) frequency modulation [11]. Thus, we have here:

$$\boldsymbol{E}_i=abs(\boldsymbol{\omega}_i-\boldsymbol{Q}_i\boldsymbol{C}_i)\ \text{ with }\ \boldsymbol{\omega}_i=\left[i\omega_0^1\ \ i\omega_0^2\ \ldots\ i\omega_0^K\right]^t \tag{7}$$

$$T_i=\left[\Delta\omega_i^1\ \ \Delta\omega_i^2\ \ldots\ \Delta\omega_i^K\right]^t\ \text{ with }\ \Delta\omega_i^k=\max(2Hz,\alpha i\omega_0^k) \tag{8}$$

where $\alpha$ is an arbitrary ratio within the range 1%–5% that controls the frequency modulation excursion, *abs* denote the element-by-element modulus function, and $\boldsymbol{Q}_i$ is the "derivate matrix” derived from $\boldsymbol{M}_i$: in each case, the general entry of $\boldsymbol{Q}_i$ is the derivative of the general entry of $\boldsymbol{M}_i$, e.g., when the polynomial model is used, the general entry of $\boldsymbol{Q}_i$ is $q_{kp}=pn_k^{p-1}$ and when the DCM is used, we have $q_{kp}=-p\pi\sin(p\pi n_k/N)/N$ .

### 3.3. Synthesis

The synthesis is achieved by applying one of equations (2–5) for phases, and a similar equation for amplitudes, depending on the chosen model, and applying (1). Since the modeling only concerns the voiced part of speech, the unvoiced sections are simply concatenated with the LT-modeled voiced sections with local overlap-add windowing to avoid audible artifacts [2].

## 4. Experiments

### 4.1. Summary of previous experiments as a reference

The perceptual criterions and the algorithm of Section 3 have been extensively assessed in [7][8], using only the (linear+) discrete-cosine LT model. The basic results are summarized here and provide a reference for the comparison between the four competing LT models that will be presented below.

A set of experiments was conducted on 8-kHz continuous speech produced by 12 different speakers (six male and six female speakers). About 3500 voiced segments of different sizes were extracted, representing more than 13 minutes of speech, and the first ten harmonics of each section were modeled.

The main results that were obtained with the (L)DCM model in [7][8] are: 1) The LT model generally fitted the amplitude or phase trajectories quite well (i.e., the modeling error is shaped just under the trajectory of the perceptual threshold model) within ten iterations of the algorithm. A value of $R_{min}$ = 0.75 was shown to be sufficient to ensure a satisfactory fitting because of the intrinsic smoothness of the model (and of the data trajectories at a lesser extend). A value of $R_{min}$ = 0.90 was also tested and allowed to increase the modeling accuracy at the price of additional iterations and increasing order. 2) Informal listening tests revealed that *the LT model generally provided a synthesis quality identical to the one obtained with usual Short-Term interpolation of the measured parameters*. This result was generally ensured with $R_{min}\geq0.75$ for both amplitudes and phases (even though the higher value of 0.9 was sometimes needed for the two first harmonics), and $\alpha\leq3\%$ for phase modeling (for $\alpha>4\%$, a difference between LT and ST synthesized signals can be heard for some speech sections). 3) The LT model was shown to provide efficient data compression: Mean coefficient rates of respectively 26 and 20.5 coefficients/s/harmonic were found for amplitudes and phases respectively, in conditions satisfying the results of 2). Comparison with the 50 frames/s of usual ST coders results in a compression factor of 2–2.5.

### 4.2. Criterions for comparison

The four competing models of Section 2 were used to model the trajectories of amplitude and phase parameters of the 3500 voiced segments of our database, by using the algorithm of Section 3. Then, we first made extensive listening tests that led to observe that the signal quality obtained with the four models (for the same conditions, $R_{min}$=0.75 or 0.9, and $\alpha$=2–4%) was identical, and identical to Short-Term synthesis for a restrained subset of these conditions (see Section 4.1). It results from this observation that other quantitative criterions must be used to compare the performances of the four models. We used the coefficient rate (defined as the average number of model coefficients necessary to model one second of speech), previously calculated for the (L)DCM model in [7][8]. We also used the percentage $P_R$ of sections for which the target ratio $R_{min}$ was reached. Indeed, the algorithm cannot guarantee that $R_{min}$ is reached for any section, since the order is limited by two factors: on the one hand, the updating process of order estimation intrinsically provides an upper limit of $K/2$, which is the limit of “over-training” the model, and on the other hand we introduced another limit resulting from computational considerations: the matrix to be inverted in (6) must not be ill-conditioned. Thus, the $P_R$ value can be seen as a quantitative measure of the “computational flexibility” of the models.

### 4.3. Results

We give on Fig. 1 the amplitude mean coefficient rates and the percentages $P_R$ obtained for the four models and the ten first harmonics, averaged over the complete corpus (for $R_{min}$=0.75). All rates are increasing with the harmonic rank. This is mainly because the complexity of amplitude trajectories generally also increase with the rank. The rates of the four models are quite close. Average values across the ten harmonics are respectively: PM: 28.2, DCM: 26.5, DCSM: 27.7 and PDCM: 28.1 coefficients/s. Thus, the DCM performs slightly better than the others (note that results are ordered slightly differently for $R_{min}$=0.90: PM: 31.5, DCM: 33.0, DCSM: 31.4 and PDCM: 31.9, but the results stay close to each other). At the same time, the $P_R$ results of the DC model outperforms the three other models, the polynomial model being the quite worse (closely followed by the PDCM). This may be because the polynomial terms are far ahead the most sensitive to computational problems, because of the large range of calculated values when the section length and/or the model order is high. On the contrary, the DCM, which is almost equivalent to the discrete cosine transform (DCT) widely-used in coding applications, is quite robust to model longer sections of speech, as the DCT is for encoding large numbers of signal samples.

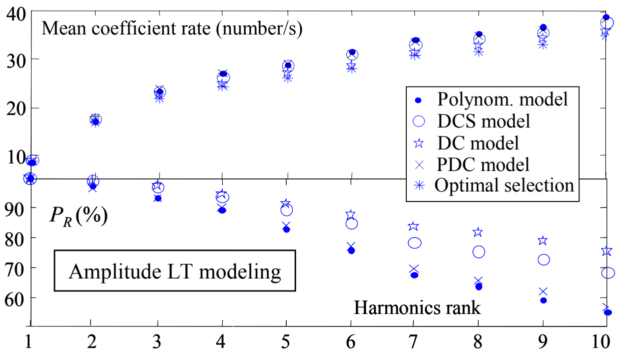

*Figure 1*: Comparative results for amplitude LT modeling; Up: Mean coefficient rates ; Down: Percentages $P_R$ (see text). Results are averaged over 3500 sections, $R_{min}$=0.75.

For the phases, the results are different (see Fig. 2). The polynomial model is the most efficient regarding the coefficient rate criterion (see Fig. 2 and the averaged values across the ten harmonics in Table 1). All values of $\alpha$ provide coherent results. According to the $P_R$ criterion, the LDCS model is first and the polynomial model is hardly lesser, but all values are high and quite close for $\alpha \geq 3\%$.

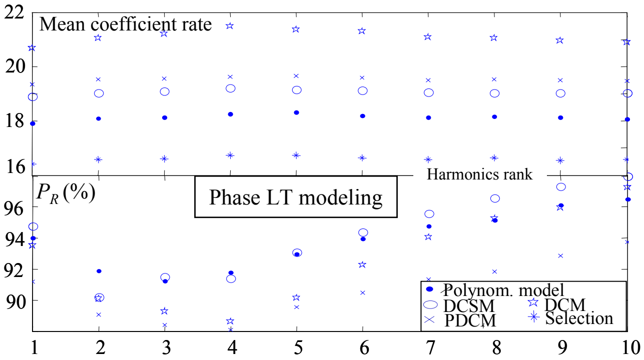

*Figure 2*: Comparative results for phase LT modeling; Up: Mean coefficient rates ; Down: Percentages $P_R$ (see text). Results are averaged over 3500 sections, $R_{min}$=0.75, $\alpha$=3%.

| model | P | DC | DCS | PDC | Selec | P | DC | DCS | PDC |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$=2% | 21.1 | 25.1 | 21.2 | 21.5 | 19.1 | 79.8 | 74.0 | 82.0 | 76.2 |
| $\alpha$=3% | 18.1 | 21.1 | 19.1 | 19.5 | 16.6 | 93.8 | 92.6 | 94.3 | 90.6 |
| $\alpha$=4% | 16.1 | 19.2 | 17.9 | 18.3 | 14.9 | 97.9 | 97.5 | 97.8 | 95.6 |

*Table 1*: Results of phase LT modeling (as those of Fig. 2) averaged across harmonics; Left: Coefficient rates; Right: $P_R$.

Given those results, it seems difficult to make a definitive choice on the "best" model for each type of parameter. Rather, the remark we previously made about the sensitivity of the polynomial model to computational problems led us to explore if each model could be "specialized" in a specific kind of voiced section, e.g. depending on its length. Thus, we plotted histograms (not shown here) describing the repartition of the "winner" model (defined as the model that needed the fewer number of parameters) for all voiced sections of the corpus. Even if the polynomial model appeared to be generally selected for the shorter voiced sections and the (L)DC model appeared to be generally selected for the longer voiced sections, the histograms were sufficiently confused so that they did not allow for a general rule for model selection. Therefore, we tested another "optimal" strategy, which consisted in simply choosing the winner model for each section. This is an optimal strategy regarding the coefficient rate criterion, but it requires to transmit additional bits to the decoder/synthesizer to encode the type of selected LT model for each section. However, since the mean number of voiced sections per second was 4.3 on our database, and 2 bits are necessary for each section to encode the information on the model type, the additional rate is very low, less than 10 bits/s, and quite lower than the bit saving corresponding to the gain on coefficients provided by optimal model selection. Indeed, the new mean coefficient rate corresponding to optimal selection among the four models, is found to be 25.7 coefficient/s for amplitude modeling with $R_{min}$=0.75 (29.7 if $R_{min}$=0.90) and respectively 19.1, 16.6, and 14.9 coefficients/s for phase modeling with $\alpha$=2%, 3% and 4% respectively ($R_{min}$=0.75) (see the "selection" values plotted on Fig. 1–2 and in Table 1). Therefore, the main new result of this study is that the use of a "multi-LT model" can save 3–10% of coefficient rate for amplitudes, and about 21–24% of coefficient rate for phases, compared to the (L)DCM alone.

## 5. Conclusion

Four different Long-Term models for sinusoidal speech parameter trajectories, based on polynomial, cosine and sine functions, were compared in terms of data compression efficiency and computational robustness, since the synthesis signal quality was found to be comparable for all models. Eventually, the selection of the "best" model for each section allow to significantly decrease the coefficient rate for both amplitudes and phases compared to the same synthesizer using only the (L)DC model (or any other of the four tested model alone). This result will be exploited in our future work, which deals with the elaboration of a very-low bit-rate "Long-Term speech coder" based on the sinusoidal model of speech, the LT modeling approach and additional quantization schemes.

## 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, 744-754.

[2] 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, 389-406.

[3] 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, 571-585.

[4] 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.

[5] R. J. McAulay & T. F. Quatieri, Sinusoidal coding, in *Speech coding and synthesis*, (W. B. Kleijn & K. Paliwal, eds), ch. 4, Elsevier, 1995.

[6] L. Girin, M. Firouzmand & S. Marchand, Long term modeling of phase trajectories within the speech sinusoidal model framework, *Proc. Int. Conf. on Speech & Lang. Proc.,* Jeju, South Korea, 2004

[7] M. Firouzmand and L. Girin, Perceptually weighted long-term modeling of sinusoidal speech amplitude trajectories, *Proc. IEEE Int. Conf. on Acoustics, Speech & Signal Proc. (ICASSP 2005),* Philadelphia, USA, 2005.

[8] L. Girin, M. Firouzmand & S. Marchand, Perceptual long-term variable-rate sinusoidal modeling of speech, *submitted to IEEE Trans. Speech and Audio Proc.*, 2005.

[9] S. Dusan, J. Flanagan, A. Karve & M. Balaraman, Speech coding using trajectory compression and multiple sensors, *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, South Korea, 2004.

[10] ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3 IT – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s, Part 3: Audio, 1992.

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