On the Estimation of the Speech Harmonic Model

Yannis Pantazis[1], Olivier Rosec[2] and Yannis Stylianou[1]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2] Orange Labs TECH/SSTP/VMI, Lannion, France
{pantazis, yannis}@csd.uoc.gr, olivier.rosec@orange-ftgroup.com

Abstract

In this paper we present and compare four time-domain approaches for estimating the parameters of a harmonic speech model. The classic approach of Least Squares is directly compared with a Total Least Squares approach trying to overcome errors in the estimation of the fundamental frequency of the model. Both of these approaches are suboptimal since they split the estimation problem into two subproblems; to the estimation of amplitudes and phases and to the estimation of fundamental frequency. To improve the accuracy of the parameters estimation of the harmonic model two iterative non linear approaches are then presented, based on the Steepest Descent and Newton-Gauss optimization algorithms, where all parameters of the harmonic model are estimated simultaneously. The approach based on the Newton-Gauss optimization algorithm provided the best accuracy as this is measured by the Signal-to-Noise Ratio criterion.

Index Terms: harmonic models, parameter estimation, speech analysis

1. Introduction

Harmonic models are able to represent efficiently various signals like speech and music. Sinusoidal models [1] may be considered as more general models than the harmonic models for speech. However harmonic models offer simplicity and efficiency in areas like speech coding [2] [3], speech synthesis and speech modification/transformation [4]. The performance of harmonic models heavily depends on the accuracy of the estimated parameters which are the fundamental frequency and the harmonic amplitudes and phases. Despite the simplicity of the harmonic model, the simultaneous estimation of all parameters of the model is not trivial since it is a nonlinear optimization problem. Since nonlinear optimization approaches are mainly iterative and therefore time consuming, they were not attractive in the past in applications like speech coding where the time delay is an important parameter in the design of the coder. However, with the increasing power of the computing systems, and since there are many other areas (i.e., speech synthesis and speech transformation) where quality of speech representation is more crucial than speed of computation, it would be interesting to investigate the accuracy of the harmonic model of speech when nonlinear optimization approaches are used for the estimation of the model parameters.

Typically, the estimation of parameters of harmonic speech models is performed into two steps; at first, the estimation of the fundamental frequency is obtained [5] [6] and then providing that the fundamental frequency is known the estimation of the harmonic amplitudes and phases is performed by minimizing a mean squared error criterion [7] [8]. Such an approach makes the parameters estimation procedure a simple linear optimization problem. Although of its simplicity, the accuracy of the estimation of the model parameters based on this approach heavily depends on the initial estimation of the fundamental frequency. This is why the estimation of the fundamental frequency has attracted the interest of many researchers in the speech analysis area.

According to the optimization theory, these linear methods produce suboptimal solutions since the estimation is decoupled into estimating the fundamental frequency first and then estimating the amplitudes and phases. In this paper, time-domain approaches for simultaneously estimation of all harmonic parameters are presented and they are compared to more traditional linear approaches. A widely used approach for computing the amplitudes and phases is through Least Squares (LS) [7], [9]. LS is equivalent to solve an overdetermined linear system and it assumes that the fundamental frequency has already been computed with sufficient accuracy. However, this is not always the case. Thus, parameter estimation through LS is vulnerable to fundamental frequency estimation. Indeed, wrong fundamental frequency estimation results in wrong amplitude and phase estimation. In this paper, different solutions for more robust amplitude and/or fundamental frequency estimation are presented. Total Least Squares (TLS) is a method that takes into account errors that may occur in the estimation of the fundamental frequency. TLS is a generalization of LS thus it is expected to provide more accurate estimates. Nonlinear Least Squares (NLS) is another solution for minimizing the sum of squared error. In NLS, fundamental frequency, amplitudes, and phases are estimated simultaneously. Given an initial estimate for the unknown parameters, iterative NLS provide fundamental frequency, amplitudes, and phases. Two different iteration schemes from optimization theory are presented. The first method is the Steepest Descent which makes use of the derivative of the error while the second method is the Newton-Gauss method where the iteration step is similar to the LS method.

Experiments were conducted on speech signals using the Harmonic+Noise (HNM) model [7]. HNM is quite efficient in representing speech signals, in performing speech modifications, and it has found to be useful in speech synthesis. In Section 2, a brief overview of HNM is provided, while Section 3 presents the Least Squares method for amplitude and phase estimation. In Section 4, the Total Least Squares method is shown. Next, Nonlinear Least Squares methods are presented in Section 5. Finally, Section 6 presents the evaluation of each estimation method on speech signals. Finally, Section 7 concludes the paper.

2. Harmonic+Noise Model

HNM decomposes speech into two components: a harmonic or deterministic component and a noise or stochastic component.

A time dependent parameter referred to as maximum voiced frequency splits the frequency axis into a lower and an upper frequency band. Harmonic component describes the lower frequency band of voiced speech segments by a sum of harmonically related sinusoids. Noise component models the upper frequency band as a time modulated colored noise [7].

Therefore, in HNM context a speech signal $s[n]$ is modeled as the sum of two components:

$$s[n] = h[n] + u[n] \ . \tag{1}$$

where $h[n]$ and $u[n]$ denote the harmonic and noise part, respectively. In this paper our focus is on the harmonic part. Harmonic part, $h[n]$, is given as:

$$h[n] = \sum_{l=-L}^{L} a_l e^{2\pi l f_0 n/f_s} \qquad n = -N, ..., N \tag{2}$$

where f_0 is the local fundamental frequency, N is the local pitch period in samples, L is the local number of harmonics and a_l are the local complex amplitudes. The number of harmonics, L, determines the order of the harmonic model. Please note that all methods presented here assume that the order of the model (i.e. the number of harmonics) is known.

The estimation of the unknown parameters of the harmonic part is done by minimizing an error criterion or cost function. The cost function is defined as the sum of squared error,

$$\epsilon(a_{-L}, ..., a_L, f_0, L) = \sum_{n=-N}^{N} (e[n])^2 \ . \tag{3}$$

The error is given by:

$$e[n] = w[n](s[n] - h[n]) \ , \tag{4}$$

with $w[n]$ being a window imposed into the error.

3. Least Squares

As it was mentioned above, LS assumes that the fundamental frequency as well as the number of harmonics are known. Then, LS minimizes the square of the error:

$$\begin{aligned} \epsilon(\mathbf{a}) &= \frac{1}{2} \sum_{n=-N}^{N} (e[n])^2 = \frac{1}{2}\mathbf{e}^h\mathbf{e} \\ &= \frac{1}{2}(\mathbf{s} - \mathbf{E}\mathbf{a})^h \mathbf{W}^2 (\mathbf{s} - \mathbf{E}\mathbf{a}) \ , \end{aligned} \tag{5}$$

where

$$\mathbf{s} = \begin{bmatrix} s[-N] & s[-N+1] & ... & s[N] \end{bmatrix}^T$$

is the vector with the speech samples,

$$\mathbf{e} = \begin{bmatrix} e[-N] & e[-N+1] & ... & e[N] \end{bmatrix}^T$$

is the vector with the error samples. The exponential matrix $\mathbf{E}$ has as elements

$$(\mathbf{E})_{N+n,L+l} = e^{j2\pi l \hat{f}_0 n/f_s}$$

with $n = -N, ..., N$ and $l = -L, ..., L$, $\mathbf{a}$ is the vector with the unknown amplitudes given by:

$$\mathbf{a} = \begin{bmatrix} a_{-L} & ... & a_0 & ... & a_L \end{bmatrix}^T$$

and $\mathbf{W}$ is a diagonal matrix with elements the samples of window $w[n]$. Note that matrix $\mathbf{E}$ has dimension $2N+1 \times 2L+1$, and the indices in this paper starts from 0. Also the fundamental frequency is denoted as $\hat{f}_0$ since it is assumed that it is arleady estimated by a pitch estimation algorithm.

LS finds the minimum of eq.(5) by setting the derivative of the error function equals to 0. Then, matrix calculus gives

$$\begin{aligned} \frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} &= -\mathbf{E}^h \mathbf{W}^2 (\mathbf{s} - \mathbf{E}\mathbf{a}) \\ &= \mathbf{E}^h \mathbf{W}^2 \mathbf{E}\mathbf{a} - \mathbf{E}^h \mathbf{W}^2 \mathbf{s} \ . \end{aligned} \tag{6}$$

Finally, setting $\frac{\partial \epsilon(\mathbf{a})}{\partial \mathbf{a}} = 0$, we obtain the LS solution (assuming the existance of the inverse which holds in this case):

$$\mathbf{a}_{LS} = (\mathbf{E}^h \mathbf{W}^2 \mathbf{E})^{-1} \mathbf{E}^h \mathbf{W}^2 \mathbf{s} \ . \tag{7}$$

4. Total Least Squares

LS assumes that errors occur only in vector s, but, as already discussed, matrix E is a function of f_0 which is not accurately estimated for every frame. Therefore, errors may occur in exponential matrix $\mathbf{E}$, too. TLS tries to minimize the sum of the squared error of $\mathbf{E}$ and $\mathbf{s}$. Solution of this problem is obtained through Singular Value Decomposition (SVD) [10].

Using SVD on matrix $\mathbf{E_s} = \mathbf{W}[\mathbf{E}|\mathbf{s}]$ we have:

$$\mathbf{E_s} = \mathbf{U}\mathbf{S}\mathbf{V}^h \ , \tag{8}$$

where columns of $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors of $\mathbf{E_s}\mathbf{E_s}^h$ and $\mathbf{E_s}^h\mathbf{E_s}$, respectively and $\mathbf{S}$ is a diagonal matrix with elements the squared eigenvalues of $\mathbf{E_s}$ sorted in decreasing order. Then, if $v_{2L+1,2L+1} \neq 0$, the solution is given as:

$$\mathbf{a}_{TLS} = -\frac{1}{v_{2L+1,2L+1}} \begin{bmatrix} v_{0,2L+1} & v_{1,2L+1} & ... & v_{2L+1,2L+1} \end{bmatrix}^T \ . \tag{9}$$

Note that when $v_{2L+1,2L+1} = 0$ then the solution is in the subspace of zero eigenvalues.

5. Nonlinear Least Squares

Using the same error criterion as in eq. (3-4) the simultaneous estimation of fundamental frequency and complex amplitudes is a nonlinear problem. Thus, there is no solution in closed form and iterative methods should be applied to solve the nonlinear least squares problem [11]. In the following we will discuss two such methods; the Steepest Descent method and the Newton-Gauss method.

5.1. Steepest Descent Method

Steepest Descent (SD) tries to find a minimum of a function given an initial guess of the location of the minimum of the function. It is a gradient type method which corrects the initial guess by moving to the opposite of the derivative (gradient) of the error function. Thus, let $\mathbf{f}(\mathbf{x})$ be a function and $\mathbf{x}^{(old)}$ be an initial guess, then $\mathbf{x}$ is updated according to

$$\mathbf{x}^{(new)} = \mathbf{x}^{(old)} - \eta \mathbf{f}^{'}(\mathbf{x}^{(old)}) \ , \tag{10}$$

where η is the rate of correction.

Under this formulation, the cost function is a function of both $\mathbf{a}$ and f_0, hence, it is written as $\epsilon(\mathbf{a}, f_0)$. Then, the partial derivative is given for f_0 by:

$$\frac{\partial \epsilon(\mathbf{a}, f_0)}{\partial f_0} = j2\pi/f_s (\mathbf{W} \cdot \mathbf{E}(\mathbf{a} \circ \mathbf{i}_L) \circ \mathbf{i}_N)^T \mathbf{e} = -\mathbf{B}\mathbf{e} \ , \tag{11}$$

where $\mathbf{i}_K = \begin{bmatrix} -K, & -K+1, & ..., & K \end{bmatrix}^T$ and $'\circ'$ denotes the Hadamard operator which means an element by element multiplication. Note that $\mathbf{B}$ is a real row vector with $2N+1$ elements. To proceed, partial derivative for amplitudes is given by:

$$\frac{\partial \epsilon(\mathbf{a}, f_0)}{\partial \mathbf{a}} = -\mathbf{E}^h \mathbf{W} \mathbf{e} \,. \tag{12}$$

Putting all this together, we obtain the iteration step:

$$\begin{bmatrix} \mathbf{a}^{(new)} \\ f_0^{(new)} \end{bmatrix} = \begin{bmatrix} \mathbf{a}^{(old)} \\ f_0^{(old)} \end{bmatrix} + \eta \begin{bmatrix} \mathbf{E}^h \mathbf{W} \\ \mathbf{B} \end{bmatrix} \mathbf{e} \,. \tag{13}$$

5.2. Newton-Gauss Method

A major difficulty of SD is that the correction rate, η, is not known. Newton-Gauss (NG) method solves this difficulty be making the correction rate an adaptive parameter which depends on the position of the estimated parameters relative to the actual minimum. Thus, NG uses the same recursion formula as SD but the correction rate is different.

More specifically, let us define matrix $\mathbf{J}$ as

$$\mathbf{J} = \begin{bmatrix} \mathbf{E}^h \mathbf{W} \\ \mathbf{B} \end{bmatrix} \tag{14}$$

i.e. the Jacobian $\mathbf{J}$ of the cost function. Then, an NG recursion leads to:

$$\begin{bmatrix} \mathbf{a}^{(new)} \\ f_0^{(new)} \end{bmatrix} = \begin{bmatrix} \mathbf{a}^{(old)} \\ f_0^{(old)} \end{bmatrix} + (\mathbf{J}^h \mathbf{J})^{-1} \mathbf{J}^h \mathbf{e} \,. \tag{15}$$

Note that the second term in (15) has a similar form as in the LS solution. However, in NG method f_0 adjustment is also performed. The convergence of the algorithm is typically 3 to 6 iterations if an initial estimate close to the true minimum is given. While each iteration for NG method is more computationally expensive than an SD iteration, the number of iterations for NG is substantially less than in SD case.

5.3. Initialization

Since SD and NG are iterative methods, initialization is necessary. Under the HNM context, we performed the following initialization. For the first frame of a voiced region, LS provides an initial estimate. For the subsequent frames then, the parameters of the previous frames are used as an initial estimate. Note that linear phase mismatch in this procedure is avoided since the analysis in HNM is performed pitch synchronously.

6. Results and Discussion

Figure 1 shows the Signal-to-Noise Ratio (SNR) in dB for a sequence of vowels ('aoie') (SIG1) uttered by a male voice while Figure 3 shows the SNR of the sentence 'vazivaza' (SIG2) which contains both vowels and voiced fricatives. The analyzed signals were sampled at 16 kHz. Furthermore, the signals have been filtered at 4 kHz and maximum voiced frequency in HNM analysis was set at 4 kHz (so no noise part is considered). In this context, "Noise" in SNR refers to modeling error. A Hamming window was used to weight the error signal since it provided the best result in [7]. The SNR difference between TLS and LS, the SNR difference between SD and LS as well as the SNR difference between NG and LS are shown in each plot of Figure 2. The corresponding SNR differences for 'vasivasa' are depicted in Figure 4.

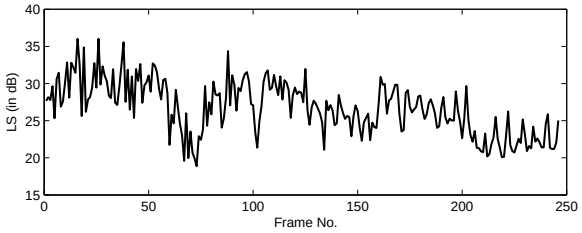

Figure 1: SNR for a sequence of vowels using the LS method.

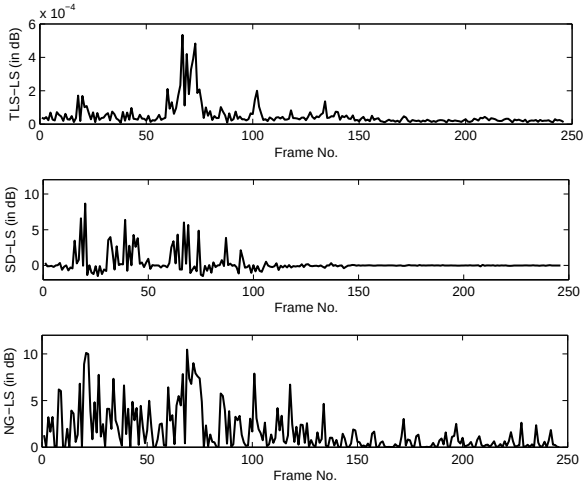

Figure 2: Signal 'aoie'. Upper plot; the SNR difference between TLS and LS. Middle plot; the SNR difference between SD and LS. Lower plot; the SNR difference between NG and LS.

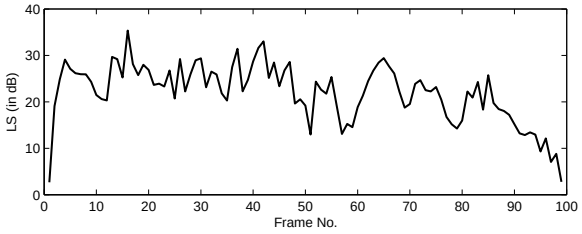

Figure 3: SNR for 'vazivaza' using the LS method.

In Table 1, the average SNR (in dB) for all the presented methods is provided. It is worth to note that NG method outperforms the other methods (in some frames there is even a near to 10 dB improvement in SNR). This is expected since the joint optimization of fundamental frequency and complex amplitudes provides solutions closer to the true minimum. SD method produced either higher and either lower SNR than LS. The reason for this behaviour is the step factor, η, which is difficult to manipulate. In this paper, correction or convergence rate, η, was set to a constant ($\eta = .001$) for each iteration and each unknown parameter. On the other hand, in NG method, the convergence rate is controlled automatically by the matrix $(\mathbf{J}^h \mathbf{J})^{-1}$ which is different in each iteration. Consequently, SD is prone to incorrect estimation because of the correction rate

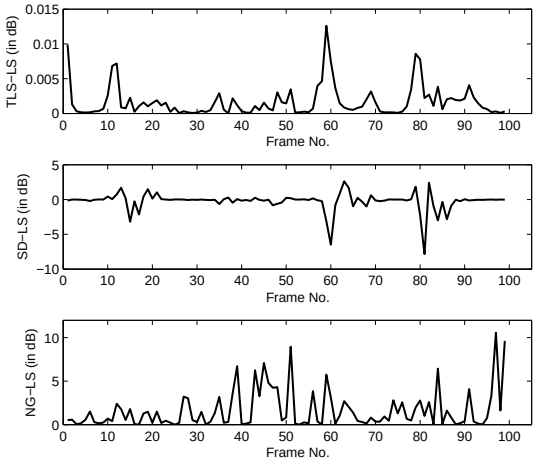

Figure 4: Signal 'vasivasa'. Upper plot; the SNR difference between TLS and LS. Middle plot; the SNR difference between SD and LS. Lower plot; the SNR difference between NG and LS..

	LS	TLS	SD	NG
SIG1	26.70	26.70	26.95	28.40
SIG2	22.05	22.05	21.80	23.70

Table 1: Average SNR in dB for the two speech examples, SIG1 and SIG2.

parameter.

NG method is not very popular due to the fact that in each iteration the inversion of $(\mathbf{J}^h\mathbf{J})^{-1}$ matrix is necessary. However, the number of iterations is quite small. In Table 2, the mean number of iterations for the nonlinear estimation methods is shown.

Regarding LS and TLS, we observe that both methods provide the same SNR. This suggests that the estimated parameters are the same. This is rather surprising since TLS is a more general method than LS. Let's try to explain this behaviour of TLS. In TLS it is assumed that the estimation of f_0 contains errors. Therefore the *true* f_0 is $f_0 + \Delta f_0 = \hat{f}_0$, where $\hat{f}_0$ is the estimated fundamental frequency. Then the element of matrix E in eq. (8) is

$$\begin{aligned}(\mathbf{E})_{N+n,L+l} &= e^{j2\pi l\hat{f}_0 n} \\ &= e^{j2\pi l(f_0+\Delta f_0)n} \\ &= e^{j2\pi l f_0 n} e^{j2\pi l \Delta f_0 n}\,.\end{aligned}$$

It turns out that the error term is inserted to the estimation problem in a multiplicative way

$$(\mathbf{E} \circ \mathbf{\Delta E})\mathbf{x} = \mathbf{s} + \mathbf{\Delta s}\,, \tag{16}$$

where $'\circ'$ as before means element by element multiplication, and not in an additive way

$$(\mathbf{E} + \mathbf{\Delta E})\mathbf{x} = \mathbf{s} + \mathbf{\Delta s} \tag{17}$$

as TLS expects. It is for this reason (multiplication instead of addition rule) that TLS is not appropriate to incorporate f_0 mismatches into harmonic models.

	SD	NG
SIG1	17.3	3.8
SIG2	10.5	5.1

Table 2: Average number of iterations for the two speech examples.

7. Conclusions

In this paper, we discussed time-domain approaches for estimating the parameters of harmonic model of speech. Both linear and nonlinear approaches were tested. Nonlinaer Least Squares using Newton-Gauss iterative method gave the highest SNR score. This can be explained from the fact that not only amplitudes and phases are recursively estimated but also the fundamental frequency. We plan to perform formal perceptual tests to further validate the performance of the discussed estimation algorithms.

8. References

[1] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34:744–754, 1986.

[2] T.F. Quatieri and R.J. McAulay. Shape-Invariant Time-Scale and Pitch Modifications of Speech. 40:497–510, 1992.

[3] Y. Agiomyrgiannakis and Y. Stylianou. The Harmonic Model Codec (HMC) Framework for VoIP. In *Interspeech*, Antwerp, Belgium, Aug 2007.

[4] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. on Speech and Audio Proc.*, 9:21–29, 2001.

[5] W. Hess. *Pitch determination of speech signals*. Springer Verlag, Berlin, 1983.

[6] Y. Stylianou J. Laroche and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.

[7] Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, 1996.

[8] Wim D'haes. *Automatic Estimation of Control Parameters for Musical Synthesis Algorithms*. PhD thesis, University of Antwerp, 2004.

[9] E. B. George and M. Smith. Analysis-by-Synthesis Overlap-Add Sinusoidal Modeling Applied to the Synthesis of Musical Tones. *Journal of the Audio Engineering Society*, 40:497–516, 1992.

[10] S. van Huffel and J. Wandewalle. *The Total Least Squares Problem; computation aspects and analysis*. Frontiers in Applied mathematics, SIAM, 1991.

[11] C. T. Kelley. *Iteative Methods for Optimization*. SIAM, 1999.