UDC 519.615.7
DOI: 10.32626/2308-5878.2018-18.135-145

V. Yu. Semenov, Ph. D.
Scientific and Production Enterprise «Delta SPE LLC», Kiev

METHOD FOR THE MAXIMIZATION OF THE LIKELIHOOD FUNCTION OF SPEECH AUTOREGRESSIVE PARAMETERS BASED ON LINE SPECTRUM PAIRS

The paper considers the estimation of the parameters of autoregressive model at additive white noise background. The principle of maximum likelihood is used for this purpose. The main goal is to find the maximum of likelihood function depending on parameters of autoregressive model. Representation of likelihood function through line spectrum pairs and other alternative parameters is presented. This provided possibility of likelihood function maximization by KNITRO algorithm. The presence of multiple local minima of the considered likelihood function is shown. Experimental results including the comparison with widely used expectation-maximization (EM) method are presented for the real speech signals.

Key words: *likelihood function, autoregressive processes, line spectrum pairs.*

Introduction. Autoregressive models are widely used in the fields of audio and image processing, analysis of economical processes and other areas. They also obtained a wide implementation in digital speech processing [1]. The majority of contemporary digital speech processing methods are based the autoregressive (AR) model of speech

$$s(l) = -\sum_{k=1}^{p} a_k s(l-k) + gw(l), \tag{1}$$

where $s(l)$ is a speech signal; $w(l)$ is an excitation process modeling the air flow at the glottis; g is a gain; $a_k, k = 1,..., p$ are the AR coefficients defining the shape of vocal tract. The order p is usually taken as 10.

When observation noise is absent, the coefficients a_k are usually found by a minimization of the criterion

$$\min_{a_1,...,a_3} \sum_{l=1}^{L} (s(l) + \sum_{k=1}^{p} a_k s(l-k))^2,$$

which leads to a system of p linear Yule-Walker equations [1].

The performance of speech processing systems based on AR model can significantly decrease at the presence of background noise, i.e. when the additive noise $v(l)$ is present and we are given only noisy observations $z(l)$:

$$z(l) = s(l) + v(l). \tag{2}$$

The degradation of such systems is explained by the lack of efficient noise-robust methods of AR parameters estimation. A big amount of works are dedicated to this subject (see, e.g. review in [2]). Starting with the paper of Lim and Oppenheim, the most popular approach to identification of AR models (1) remains maximization of likelihood function (LF) of speech AR parameters [2]. A majority of existing methods are the modifications of Expectation-Maximization (EM) method which provides maximum likelihood estimation of speech AR parameters. The rigorous substantiation and implementation of this method was given in work of Gannot *et al.* [2] which can be considered as a reference point for other methods. The weak points of this approach are the necessity of good initial approximation and high computational expenses of Kalman smoothing.

For the maximization of likelihood function, we derive its alternative representations using auxiliary parameters such as line spectrum pairs (see, e.g. [3]). For the optimization of simplified likelihood function we employ KNITRO optimization algorithm which is now widely used for global optimization tasks [4, 5].

The structure of the paper is as follows. First, we define the likelihood function and its expression via the line spectrum pairs. We then show that the likelihood function of speech AR parameters may have multiple local minima. After introducing one more alternative representation of likelihood function, we show the results of implementation of proposed approach to the estimation of AR parameters of real speech signals.

Likelihood function of AR parameters. Consider speech frame which is a vector of L observations $Z = [z(1),...,z(L)]$, where the values $z(l)$ are given by (2). The goal is to estimate AR coefficients a_k ($k = 1,..., p$) and the gain g having the vector Z.

In speech coding applications, L is usually can taken as 200, which is equivalent to 25 ms when the sampling frequency is 8000 Hz (it is usually supposed that the parameters a_k ($k = 1,..., p$) and g can be treated as constant during approximately 25 ms). In the following, we use values $p = 10$ and $L = 200$.

The likelihood function of observation vector can be presented as:

$$f(Z \mid a_1,..., a_p, g) = \frac{1}{\sqrt{\det 2\pi C}} \exp(-\frac{1}{2} Z^T C^{-1} Z), \tag{3}$$

which is a conditional Gaussian probability density of the observations vector Z. Here C denotes the covariance matrix of Z; it depends on the gain and the AR coefficients. The principle of maximum likelihood states that given Z, the optimal parameters maximize the likelihood function (3) subject to some physical constraints.

Line spectrum pairs formulation. The direct computation of (3) based on the models (1) and (2) is complicated [2, 6]. A better representation based on the spectrum $P_s(j)$ of the AR process $s(n)$ is given by [6]:

$$P_s(j) = \frac{g^2}{|1+\sum_{k=1}^{p} a_k e^{-2\pi ikj/L}|^2} \quad (j=1,...,L). \tag{4}$$

In terms of the spectrum, we may represent the negative logarithm of the likelihood as proposed in [6]:

$$f(Z\,|\,a_1,a_2,...,a_p,g) = \sum_{j=1}^{L} \log(P_s(j)+P(j)) + \sum_{j=1}^{L} \frac{|Z(j)|^2}{P_s(j)+P(j)}. \tag{5}$$

Here the $Z(j)$ form the spectrum of the observations, i.e., the discrete Fourier transform of Z ; $P(j)$ is the known spectrum of the observational noise $v(l)$, estimated from the frames which do not contain speech.

In order to incorporate the stability constraint, we use the spectral representation defined in [7], according to which the AR spectrum (4) can be written in the form

$$P_s(j) = \frac{\xi}{F_j(x)} \quad (j=1,...,L), \tag{6}$$

where $\xi = g^2/2^{p-1}$, $x=(x_1,x_2,...,x_p)$ is the vector of ordered cosines of line spectrum pairs (LSP) $x_k = \cos(\omega_{p-k})$ $(k=1,2,...,p),$ and

$$F_j(x) = (1-c_j)(\prod_{k\le p\ \ odd}(c_j-x_k))^2 + (1+c_j)(\prod_{k\le p\ \ even}(c_j-x_k))^2 \tag{7}$$

with the constants $c_j = \cos(2\pi j/L)$. The parameters x_k are ordered so that they satisfy the constraint

$$-1 < x_1 < x_2 < x_3 < ... < x_{p-1} < x_p < 1\,. \tag{8}$$

Thus, the spectrum (6) is determined by the spectral parameters $x_1,x_2,...,x_p$ and the gain ξ .

Simplification of the optimization problem. To simplify the statement of the problem, first, we introduce the constants

$$Z_j = |Z(j)|^2$$

and the functions

$$Q_j(x,\xi) = F_j(x)/[P_jF_j(x)+\xi]. \tag{9}$$

So, we have to minimize the function

$$f(x_1, x_2, ..., x_p, \xi) = \sum_{j=1}^{L} (-\log Q_j(x, \xi) + Z_j Q_j(x, \xi)), \qquad (10)$$

with $Q_j(x)$ is defined by (9).

Our goal is to find the minimum of the function (10), taking into account the constraints (8) and $\xi > 0$. For this purpose, we do the transformations (simplifications) of the problem and apply the KNITRO [4, 5] optimization algorithm.

Multiple local minima. Consider the AR signal $s(l)$ generated by the model (1) with $g = 0.388$, white noise excitation $w(l)$ and AR coefficients

$a = [-1.195, 0.624, 0.017, -0.345, 0.259, 0.121, -0.239, 0.348, -0.210, -0.098]$.

(the coefficients correspond to a vowel sound «a» pronounced by a male speaker). These coefficients correspond to $\xi = 2.95 \times 10^{-4}$ and line spectrum cosines

$x_0 = [-0.941, -0.834, -0.473, -0.168, 0.092, 0.360, 0.493, 0.825, 0.878, 0.963]$.

The signal $s(l)$ is then mixed according to the observation model (2) with white noise $v(l)$ with constant spectrum $P_j = 0.194$ ($j = 1, ..., L$), corresponding to a signal-to-noise ratio of $SNR = 5$ dB.

The KNITRO solver, started from different initial approximations, found seven local minima of the function (10) with the constraint (8). These points are presented in Table 1. Note that minima 3, 4, 7 are *degenerate*, i.e., have $x_k = x_{k+1}$ for some k. The minima are sorted by increasing of objective function values.

The last two rows of Table 1 present the spectral distortion (SD) and Itakura-Saito measures (IS) between the ideal parameters (x_0, ξ_0) and the parameters (x, ξ) of corresponding minimum. These measures are commonly used in the analysis of speech AR parameters (see, e.g. [6]). In our terms of F, ξ these measures can be expressed as follows

$$SD(x_1, x_2) = \sqrt{\frac{1}{L} \sum_{j=1}^{L} \left[10 \log_{10} \frac{F_j(x_1)}{F_j(x_2)} \right]^2},$$

(note that the SD is usually calculated as gain-independent [6]; hence there is no factor ξ_1 / ξ_2).

$$IS(x_1, \xi_1, x_2, \xi_2) = \frac{1}{L} \sum_{j=1}^{L} \left[\frac{\xi_1}{\xi_2} \frac{F_j(x_2)}{F_j(x_1)} - \log\left(\frac{\xi_1}{\xi_2} \frac{F_j(x_2)}{F_j(x_1)}\right) - 1 \right].$$

The data in Table 1 show that the global minimizer fits the best by both criteria.

Table 1

Several local minima of function (10)

	Min.1	**Min.2**	**Min.3**	**Min.4**	**Min.5**	**Min.6**	**Min.7**
x_1	–0.991	–0.499	–0.994	–0.235	–0.997	–0.988	–0.908
x_2	–0.499	–0.468	–0.469	0.009	–0.294	–0.977	–0.908
x_3	–0.453	–0.347	–0.469	0.009	–0.187	0.026	–0.565
x_4	–0.116	–0.012	–0.016	0.332	–0.033	0.399	0.299
x_5	0.076	0.144	0.236	0.45	0.22	0.460	0.422
x_6	0.433	0.499	0.47	0.613	0.486	0.626	0.565
x_7	0.622	0.647	0.647	0.744	0.653	0.739	0.709
x_8	0.829	0.841	0.836	0.861	0.84	0.860	0.852
x_9	0.875	0.878	0.877	0.888	0.878	0.886	0.883
x_{10}	0.955	0.962	0.959	0.978	0.962	0.977	0.97
ξ	1.1×10^{-4}	4.0×10^{-5}	6.2×10^{-5}	1.7×10^{-6}	3.9×10^{-5}	4.7×10^{-6}	1.7×10^{-5}
f	24.224	24.352	24.445	25.785	26.004	28.96	29.134
$\lVert x-x_0\rVert$	0.37	0.64	0.46	1.39	0.67	0.93	0.66
SD	7.0	12.1	8.3	22.2	10.1	15.8	10.7
IS	7.4	221.0	23.4	6.5×10^{4}	78.9	1.2×10^{3}	124.3

Alternative formulation. Let's introduce another set of parameters related with LSP. This formulation is mathematically identical to our initial formulation, however:

(i) The formulation is based on more physical quantities. Unlike $\log F_j$, the quantities $\log G_j = -\log P_s(j)$ have a direct physical interpretation.

(ii) In particular, positive lower bounds on the new variables x'_j imply the nondegeneracy of the spectrum.

(iii) The interval analysis may also improve since the dependence structure is different.

Thus, we introduce the new set of variables

$$z_{p+1} := \xi^{-1/p}, \quad z_i := z_{p+1}x_i \quad (i = 1,\dots, p),$$

so that

$$x_i = z_i / z_{p+1}, \quad \xi = z_{p+1}^{-p}.$$

Then

$$G_j(z) = P_s(j)^{-1} = F_j / \xi$$

is still polynomial:

$$G_j(z) = (1-c_j)\left(\prod_{k\le p\ \ odd}(z_{p+1}c_j - z_k)\right)^2 + \left(1+c_j)(\prod_{k\le p\ \ even}(z_{p+1}c_j - z_k)\right)^2.$$

Besides,

$$Q_j(x,\xi) = R_j(x,\xi)^{-1},$$

where

$$R_j(x,\xi) := P_j + G_j(z)^{-1}.$$

Therefore, the objective function (10) becomes

$$f(z) = \sum_j(\log R_j + \frac{Z_j}{R_j}),$$

and the constraint (8) is now

$$-z_{p+1} = z_0 < z_1 < ... < z_p < z_{p+1}.$$

By introducing the positive variables

$$x_i' = z_i - z_{i-1} \quad (i = 1,2,..., p+1),$$

we get

$$z_i = \frac{1}{2}\sum_j \varepsilon_{ij}x_j',$$

where $\varepsilon_{ij} = 1$ if $j \le i$ and $\varepsilon_{ij} = -1$ otherwise.

The a priori ranges for $x_j', j = 1,..., p+1$, estimated from clean speech data, are given in Table 2 (to compute statistical quantities representative for speech, we used $N = 23294$ frames of clean speech).

Table 2

Bounds for the variables x_j'

	x_1'	x_2'	x_3'	x_4'	x_5'	x_6'	x_7'	x_8'	x_9'	x_{10}'	x_{11}'
Lower	0.02	0.03	0.03	0.049	0.080	0.064	0.074	0.049	0.033	0.006	0.002
Upper	0.869	0.972	1.860	2.128	2.377	2.392	2.444	2.025	1.542	0.784	0.209

Thus we have a bound-constrained problem in the x_j'. We need to minimize function

$$f_{\ alt}(x') = \sum_j(\log R_j(x') + \frac{Z_j}{R_j(x')}) \qquad (11)$$

where

$$R_j(x') = P_j + G_j(x')^{-1}$$

and $G_j(x')$ is given by

$$G_j(x') = 2^{-p}[(1-c_j)(\prod_{k\le p \;\; odd}[(c_j-1)\sum_{j=1}^{k} x'_k + (c_j+1)\sum_{j=k+1}^{p+1} x'_k])^2 +$$

$$(1+c_j)(\prod_{k\le p \;\; even}[(c_j-1)\sum_{j=1}^{k} x'_k + (c_j+1)\sum_{j=k+1}^{p+1} x'_k])^2].$$

This optimization problem is subject to the constraints

$$\underline{x'} \le x' \le \overline{x'} \tag{12}$$

with the bounds as defined in Table 2.

Multiple local minima (alternative formulation). Consider the example from Section 3.2 in x'-domain. The parameters $\xi = 2.95\times 10^{-4}$ and x_0 correspond to

$x'_0 = [0.132, 0.243, 0.812, 0.689, 0.586, 0.603, 0.300, 0.750, 0.119, 0.192, 0.083]$ (this point satisfies the constraint (12)).

The KNITRO solver, started from different initial approximations, found three local minima of the function (11) with the constraint (12). To compare with Table 1, we transformed these points to x-domain. They are presented in Table 3.

Table 3

Minima 1–3 of function (11) with constraint (12)

	Min.1	Min.2	Min.3
x_1	–0.989	–0.991	–0.871
x_2	–0.580	–0.692	–0.858
x_3	–0.457	–0.670	–0.504
x_4	–0.196	0.069	0.160
x_5	0.027	0.261	0.281
x_6	0.403	0.480	0.516
x_7	0.600	0.654	0.667
x_8	0.824	0.837	0.843
x_9	0.874	0.878	0.879
x_{10}	0.951	0.960	0.964
ξ	1.7e–4	6.7e–5	4.1e–5
f	25.17	28.03	29.38
$\left\| x - x_0 \right\|$	0.3	0.4	0.5
SD	5.5	7.4	7.9
IS	2.4	11.9	22.1

Comparing Tables 1 and 3, we see that the «alternative» global minimizer provides better values of quality criterions (especially Itakura-Saito measure). The G-plots are present at the Figure 1. All minima quite well approximate the clean spectrum at the lower frequencies while the global one gives the closest approximation for the middle and upper frequencies.

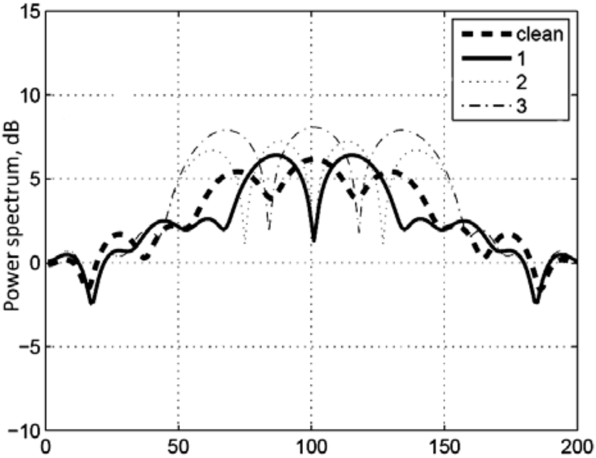

Fig. 1. *Spectra* $\log G_j$ *(*$j = 1,...,L$*) for the clean spectrum and three local minima of the function (11)*

Here arises a question: why there are just three local minima? What about the other minima from Table 1? The answer is that those minima does not satisfy constraint (12). From the previous experience, additional minima may arise due to the introduction of additional constraints.

It was also noted that for the alternative formulation the KNITRO algorithm works faster (i.e. it required lesser number of iterations) as compared with initial one (probably, since the problem is bound constrained).

Experimental results. Consider a fragment of real speech signal pronounced by male speaker. The duration of fragment is 40 frames (8000 samples, i.e., 1 sec). White noise with signal-to-noise ratio of 5 dB was used. We compared KNITRO with widely used Expectation-Maximization (EM) method which provides the local maximization of likelihood function [2, 6]. The application of EM method to the estimation of speech AR parameters in frequency domain was shown in [3]. It does not require the computation of derivatives and has a simple realization. The KNITRO algorithm was initialized by the estimate from the previous frame. The constraints (12) were used. The number of iterations was limited to 25. The first results showed the effect of the «narrow peaks». The typical situation is shown at the Figure 2.

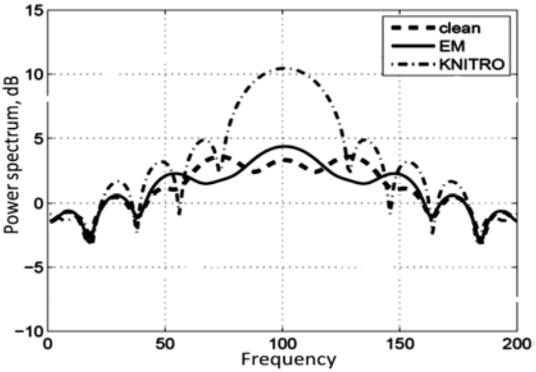

Fig. 2. *Speech example 1: Spectra* $\log G_j$ *(* $j = 1,...,L$ *) for the clean spectrum, EM spectrum and KNITRO spectrum for the real speech frame*

Since the narrow peaks are explained by the small differences of variables x_i', we shortened the range used in (12). The 90% percentile limits are shown in Table 4. It allowed to improve the situation. The example is shown at the Figure 3 (all the data are the same as for the Figure 2). It can be seen that speech resonances at the lower frequencies are now reproduced much more accurately.

The average values of Spectral Distortion and Itakura-Saito measures for 40 frames are given in Table 5.

Table 4

Stricter bounds for the variables $x_{j'}$ *(90% percentile limits)*

	$x_{1'}$	$x_{2'}$	$x_{3'}$	$x_{4'}$	$x_{5'}$	$x_{6'}$	$x_{7'}$	$x_{8'}$	$x_{9'}$	$x_{10'}$	$x_{11'}$
Lower	0.15	0.16	0.24	0.29	0.46	0.45	0.41	0.35	0.11	0.02	0.02
Upper	0.45	0.61	0.72	1.09	1.36	1.36	1.25	1.05	0.61	0.23	0.09

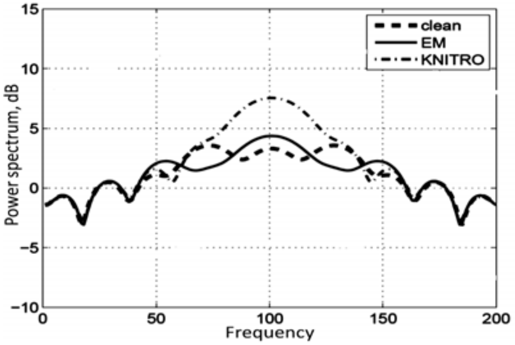

Fig. 3. *Speech example 1: Spectra* $\log G_j$ *(* $j = 1,...,L$ *) for the clean spectrum, EM spectrum and KNITRO spectrum for the real speech frame with 90%-constraints on* x_j'

Table 5

Speech quality criteria for the EM and KNITRO methods

	EM	**KNITRO**
SD	5.5	5.8
IS	14.7	5.4

The results from Table 5 show that both EM and KNITRO algoritms provide comparable mean spectral distorition value (5.5 and 5.8 respectively), while the value of Itakura-Saito mesure is much less for the KNITRO algorithm. This can be explained by more precise estimation of gain coefficient by proposed approach while the general form of spectrum for both methods is approximately the same.

Conclusions. In this paper we considered estimation of parameters of autoregressive model at noise background. At first, we introduced representation of likelihood function via line spectrum pairs and additional equivalent set of parameters. We have shown the presence of multiple local minima of likelihood function for speech autoregressive parameters. For the optimization of objective function the KNITRO algorithm was implemented. The preliminary experimental results show that the proposed approach provides better performance in comparison with widely used expectation-maximization (EM) in terms of Itakura-Saito measure while mean spectral distortion value is approximately the same. This can be explained by more precise estimation of gain coefficient by proposed approach while the general form of spectrum for both methods is approximately the same.

Acknoledgement. The author expresses gratitude to professor Arnold Neumaier from the University of Vienna for many fruitful suggestions during this investigation.

References:

1. Rabiner L. R. Theory and Applications of Digital Speech Processing / L. R. Rabiner, R. W. Schafer. — N.-J. : Prentice-Hall, 2011.
2. Gannot S. Iterative and sequential Kalman filter-based speech enhancement algorithms / S. Gannot, D. Burnstein, E. Weinstein // Transactions Speech Audio Processing. — 1998. — Vol. 6. — P. 373–385.
3. Semenov V. A novel approach to calculation of line spectral frequencies based on inter-frame ordering property / V. Semenov // Proc. IEEE Conf. ICASSP. — 1998. — Vol. 6. — P. 1072–1075.
4. Domes F. The optimization test environment / F. Domes, M. Fuchs, H. Schichl, A. Neumaier // Optimization and Engineering. — 2014. — Vol. 15. — P. 443–468.
5. Byrd R. Large-Scale Nonlinear Optimization, chapter KNITRO: An Integrated Package for Nonlinear Optimization / R. Byrd, J. Nocedal, R. Waltz. — Springer, 2006. — P. 35–59.

6. Kalyuzhny A. Ya. A method for identification of speech autoregressive parameters in frequency domain / A. Ya. Kalyuzhny, A. A. Kovtonyuk, V. Yu. Semenov // Acoustic bulletin. — 2010. — Vol. 13. — №2. — P. 20–27.
7. McLoughlin I. V. Line Spectral Pairs / I. V. McLoughlin // Signal Proces. — 2008. —Vol. 88. — P. 448–467.

МЕТОД МАКСИМІЗАЦІЇ ФУНКЦІЇ ПРАВДОПОДІБНОСТІ АВТОРЕГРЕСИВНИХ ПАРАМЕТРІВ МОВНОГО СИГНАЛУ, ЗАСНОВАНИЙ НА ВИКОРИСТАННІ ЛІНІЙНИХ СПЕКТРАЛЬНИХ ПАР

У статті розглянуто задачу оцінювання параметрів авторегресивної моделі за наявності адитивного білого шуму. Для цього застосовано принцип максимальної правдоподібності. Головним завданням є знаходження глобального максимуму функції правдоподібності, що залежить від параметрів авторегресивної моделі. Отримано вираз функції правдоподібності через лінійні спектральні пари та інший альтернативний набір параметрів. Це надало можливість ефективної максимізації функції правдоподібності за допомогою алгоритму KNITRO. Показано наявність багатьох локальних мінімумів функції правдоподібності для авторегресивних параметрів мовних сигналів. Також представлені експериментальні результати, що включають порівняння із загальновикористовуваним методом expectation-maximization (EM) для реальних мовних сигналів.

Ключові слова: *функція правдоподібності, авторегресивний процес, лінійні спектральні пари.*

Отримано: 23.11.2018