ON ROBUST TIME-VARYING AR SPEECH ANALYSIS BASED ON T-DISTRIBUTION

Dejan Bajić
Institute of Applied Mathematics and Electronics,
Kneza Miloša 37, 11000 Belgrade, Yugoslavia
fax: +381 11 186-105, e-mail: EBAJICD@UBBG.ETF.BG.AC.YU

ABSTRACT

In this paper a new robust non-recursive algorithm for parameter estimation of AR model of speech signal is proposed. The proposed algorithm takes into account the quasi-periodic excitation for voiced speech and assumes the t-distribution with small degrees of freedom α of the excitation signal. The method is based on the covariance linear prediction with sliding window. Experiments on both synthesized and natural speeches have shown that the proposed robust algorithm gives estimates with smaller variance and bias, compared to the conventional non-robust algorithm. The choice of α=3 induces to the most efficient estimation.

1. INTRODUCTION

In the conventional linear prediction speech analysis the LP parameters are determined by either the auto-correlation method or the covariance method (CLP) [1]. CLP procedure minimize the sum of squared residuals and weight all prediction residuals equally. Therefore, the result is a least squares (LS) type algorithm. The LS method can achieve good estimation results only when the driving source is a Gaussian process. It is well known that in many cases the source is of a quasi-periodic nature with spiky excitation that is not a Gaussian process, such as voiced-speech signals. For these kinds of processes, the obtained results from the LS type algorithms are biased and inefficient. The obtained estimates are very much affected by strong signal parts. We cannot accurately and efficiently estimate the parameters of the process.

In this paper, a new robust algorithm for the linear prediction (RBLP) analysis of speech is considered. The method is based on the t-distribution and covariance LP method with sliding window. In the proposed method we use a loss function that assigns large weighting factor for small amplitude residuals and small weighting factor for large amplitude residuals that are caused by the pitch excitations. The loss function is based on the assumption that the residual signal has an independent and identical t-distribution t(α) with α degrees of freedom [2]. The efficiency of this new estimator depends on α. When $\alpha \to \infty$, we get the CLP method. When the proposed method with small α=3 is applied to the problems of estimating the parameters of AR model of the synthetic speech, we can achieve a more accurate estimate and a smaller standard deviation (SD) than that with large α. The t-distribution with small α has more probability on its tail than that with large α. By using t-distribution with small α assumption, we assume that the residual signal is more spiky than in Gaussian assumptions. By using small α, a better separation between the source excitation and the vocal tract system can be achieved.

The paper is organized as follows: description of the proposed method is given in Section 2, experimental analysis is presented in Section 3, conclusions are provided in Section 4.

2. ROBUST PROCEDURE

The residual signal ε_k can be expressed as a function of the linear prediction (LP) vector as

$$\varepsilon_i(a) = s_i + \sum_{j=1}^{p} a_j \cdot s_{i-j} \qquad (2.1)$$

where $a = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}^T$; a_i are LP coefficients.

The speech signal s_i is observed along a window; $1 \le i \le M$, M is number of samples. When the exciting distribution differs from Gaussian, least-squares criteria leads to wrong results. Thus we minimized the sum of nonlinear residual function, instead of minimizing the squared residuals sum. We choose nonlinearity on a such way, that its appliance down-weights influence of a small number of large residuals.

$$J_M(a) = \sum_{i=1}^{M} \rho\left[\varepsilon_i(a)\right] \qquad (2.2)$$

Loss function $\rho(z)$ in (2.2) can be chosen as $\rho(z) = -\log f(z|a)$ and the solution of (2.2) is the maximum likelihood (ML) estimate. $f(z|a)$ is probability density function (PDF). The residual signal is assumed to have an independent and identical distribution (IID) $f(x)$.

The logarithmic of the residual likelihood function is

$$L(a|\varepsilon) = \log \prod_{i=p+1}^{M} f\left(\varepsilon_i(a)\right) = \sum_{i=p+1}^{M} \log f\left(\varepsilon_i(a)\right) \qquad (2.3)$$

where $\varepsilon = \begin{bmatrix} \varepsilon_{p+1} & \varepsilon_{p+2} & \cdots & \varepsilon_M \end{bmatrix}^T$

The loss function is $\log f\left(\varepsilon_i(a)\right)$ and the influence function is defined as

$$\beta(x) = -\frac{\partial \log f(x)}{\partial x} \qquad (2.4)$$

The Gaussian distribution

$$f_G(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2} \qquad (2.5)$$

is used for $f(x)$ in the CLP speech analysis.
The Huber's probability density function

$$f_H(x) = \frac{1-\theta}{\sqrt{2\pi}} \exp^{-\rho_H(x)} \qquad (2.6)$$

$$\rho_H(x) = \begin{cases} c|x| - \dfrac{c^2}{2} & \text{za } |x| \geq c \\ \dfrac{x^2}{2} & \text{za } |x| \leq c \end{cases} \qquad (2.7)$$

is used as $f(x)$ in the Huber's M-estimation. For heavy-tailed distribution processes the Huber's M-estimate is more efficient than the CLP method. This is because the Huber distribution is heavy-tailed, so that the influence function $\beta(x)$ assigns less weight for the large residuals caused by the spiky excitation. In this paper we proposed to use the heavy-tailed t-distribution model to construct an M-estimate.

The t-distribution with α degrees of freedom, t(α) is defined by [2]

$$f_\alpha(x) = \frac{1}{\sqrt{\alpha\pi}} \frac{\Gamma\left(\dfrac{\alpha+1}{2}\right)}{\Gamma\left(\dfrac{\alpha}{2}\right)} \frac{1}{\left(1+\dfrac{x^2}{\alpha}\right)^{(\alpha+1)/2}} \qquad (2.8)$$

For $t(\infty)$, f_α is the Gaussian distribution with zero mean and standard deviation (SD) equal to one. For the estimation purpose, $f(x)$ has to have a finite second moment. $f(x)$ has an infinite second moment for $\alpha<3$ hence, we use $\alpha \geq 3$. Research work showed that choice of small degrees of freedom $\alpha=3$ induces to the most accurate and the efficient estimation. Optimal linear prediction vector a is selected by maximizing the likelihood residual function in Eq. (2.3). The loss function is modified in the similar manner as the covariance method.

$$L(a|\varepsilon) = K_\alpha - e\hat{L}(a) \qquad (2.9)$$

where $\varepsilon = \begin{bmatrix} \varepsilon_{p+1} & \varepsilon_{p+2} & \cdots & \varepsilon_M \end{bmatrix}^T$

$$K_\alpha = (M-p)\log\left(\frac{1}{\sqrt{\alpha\pi}} \frac{\Gamma\left(\dfrac{\alpha+1}{2}\right)}{\Gamma\left(\dfrac{\alpha}{2}\right)}\right) \qquad (2.10)$$

$$e = \frac{\alpha+1}{2} \qquad (2.11)$$

$$\hat{L}(a) = \sum_{i=p+1}^{M} \log\left(1 + \frac{\left(\dfrac{\varepsilon_i(a)}{\hat{s}}\right)^2}{\alpha}\right) \qquad (2.12)$$

To get a scale-invariant estimate, residual ε_i is normalized with $\hat{s}$. The factor $\hat{s}$ in (2.13) provides to obtain a scale-invariant version of the estimator. In this work, we used robust estimate of $\hat{s}$ defined as:

$$\hat{s} = \text{median}|\varepsilon_i|, \quad p+1 \leq i \leq M \qquad (2.13)$$

Maximizing $L(a|\varepsilon)$ in (2.9) is equivalent to minimizing $\hat{L}(a)$ in (2.12), because K_α and e are both constants. The observed relation between the loss function and the desired AR coefficient is nonlinear. Therefore a Newton-Raphson iterative method need to be used to obtain the optimal coefficients set a:

$$Ga^{k+1} = Ga^k - \nabla, \qquad (2.14)$$

where k is iteration number, ∇ is gradient vector

$$\nabla = \begin{bmatrix} \dfrac{\partial\hat{L}(a)}{\partial a_1} & \dfrac{\partial\hat{L}(a)}{\partial a_2} & \cdots & \dfrac{\partial\hat{L}(a)}{\partial a_p} \end{bmatrix}^T \qquad (2.15)$$

where

$$\frac{\partial\hat{L}(a)}{\partial a_z} = \frac{2}{\alpha\hat{s}^2} \sum_{i=1}^{M+p} \varepsilon_i \cdot w_i \cdot s_{i-z} \qquad (2.16)$$

$$w_i = \frac{1}{1 + \dfrac{(\varepsilon_i/\hat{s})^2}{\alpha}} \qquad (2.17)$$

$$G = \begin{bmatrix} G_{1,1} & G_{1,2} & \cdots & G_{1,p} \\ G_{2,1} & G_{2,2} & \cdots & G_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ G_{p,1} & G_{p,2} & \cdots & G_{p,p} \end{bmatrix}, \qquad (2.18)$$

$$G_{r,t} = G_{t,r} = \frac{2}{\alpha\hat{s}^2} \sum_{i=1}^{M+p} s_{i-r} \cdot s_{i-t} \cdot w_i \qquad (2.19)$$

The result from the CLP is used as the starting value. Criteria (2.20) and (2.21) are used to terminate the iteration. It is shown by simulation results that only few iterations were need to reach a stationary point and that 10^{-4} was a suitable value for stopping the iteration. No further significant improvements can be achieved when a value lower than 10^{-4} is used.

$$\sqrt{\sum_{z=1}^{p} \left(\frac{\partial\hat{L}(a^k)}{\partial a_z}\right)^2} \leq 10^{-4} \qquad (2.20)$$

$$\left|\hat{L}(a^k) - L(a^{k-1})\right| \leq 10^{-4} \qquad (2.21)$$

The proposed algorithm can be summarized as follows:

1. Calculate the initial a by the CLP method
2. Calculate new $\hat{s}$ based on Eq (2.13)
3. Calculate new a based on Eq. (2.14)
4. Repeat step 2 and 3 until either one or both of the stopping criteria in Eq(2.20) and (2.21) are reached.

3. EXPERIMENTAL ANALYSIS

The proposed RBLP algorithm have been tested on both synthetic and natural speeches. The length of the sliding window was 256 samples, while the sliding window step was equal to one. For purposes of comparison, the synthetic and separately spoken vowels were used.

3.1. Testing on synthetic data obtained by filtering the excitation pulse train

To compare the performance of the algorithms, the test signal for the vowel [a] was synthesized by filtering an excitation pulse train with the pitch period T_p=8ms. The following formant center frequencies F_i and their bandwidths B_i, i=1,...,4, in the spectral domain were used for the vowel [a]: F_1=730 Hz, B_1=60 Hz, F_2=1090 Hz, B_2=100 Hz, F_3=2440 Hz, B_3=120 Hz, F_4=3500 Hz, B_4=175 Hz. Assuming a sampling rate of 10 kHz, in the discrete time domain this corresponds to the eight-order AR model with the following parameters: AR_1=-2.221, AR_2=2.895, AR_3=-3.088, AR_4=3.277, AR_5=-2.774, AR_6=2.355, AR_7=-1.67, AR_8=0.751. The trajectories of AR parameters obtained by the standard CLP method and RBLP method is presented in Fig. 1 and Fig. 2 respectively. In the Fig. 3 we can clearly see the influence of the Dirak pulses with period 2.5ms to the bias and variation of CLP estimated trajectories. These influences were absolutely suppressed by RBLP method.

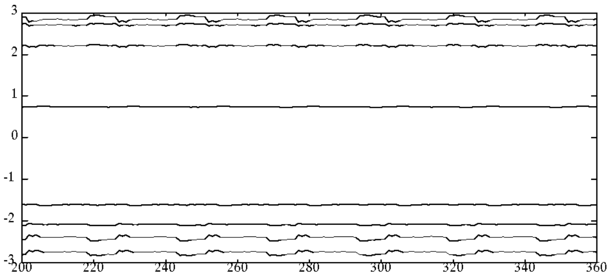

Figure 1: AR parameters obtained by CLP

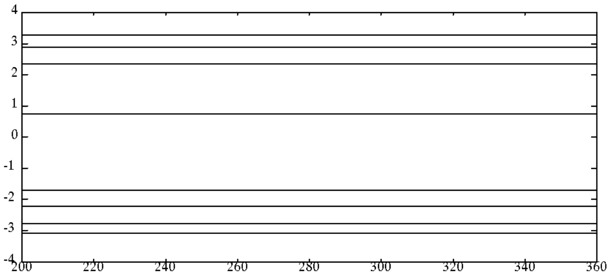

Figure 2: AR parameters obtained by RBLP

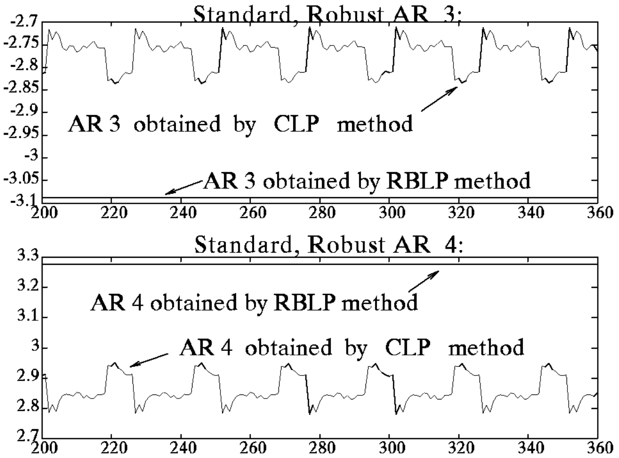

Figure 3. Zoomed AR_3 and AR_4 parameters

Table 1.	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	AR_7	AR_8
Aps. err.	0.131	0.178	0.316	0.415	0.361	0.143	0.086	0.009
Rel. err.	0.059	0.061	0.102	0.127	0.130	0.061	0.051	0.012
Std. dev.	0.012	0.019	0.037	0.048	0.039	0.015	0.011	0.006

Table 1: Standard CLP method

Table 2.	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	AR_7	AR_8	
Aps. err.	0.39	1.04	1.75	2.18	2.07	1.49	0.84	0.27	$\cdot 10^{-13}$
Rel. err.	1.74	3.59	5.65	6.65	7.45	6.32	4.93	3.64	$\cdot 10^{-14}$
Std. dev.	0.49	1.33	2.22	2.76	2.61	1.86	1.04	0.34	$\cdot 10^{-13}$

Table 2: RBLP method

Statistical data of trajectories obtained by CLP method is presented in the table 1, while the table 2 contains data about the robust procedure. Absolute and relative mean-errors are calculated, as well as standard deviations. Tables show the superiority of the robust procedure, which is reflected in much smaller value of variance and smaller bias of parameter estimates.

3.2. Testing on natural speech data

In the case of natural human speech, the true values of the vocal tract parameters are unknown. The AR parameter estimates obtained on sliding window shorter than the pitch period was used as the reference trajectory [3]. The experimental analysis was performed on isolated vowels, filtered by a low-pass filter with an upper limit frequency F_g=4kHz, and digitized by a 12-bit A/D conversion with a sampling rate of 10 kHz. In addition, preemphasis of the speech signal was also performed. In all the experiments, the AR model of the 8th order is used, and the results of the estimation of the a1-a8 AR parameters for the vowel 'a' are presented in Fig. 4. Here the dotted lines represent reference

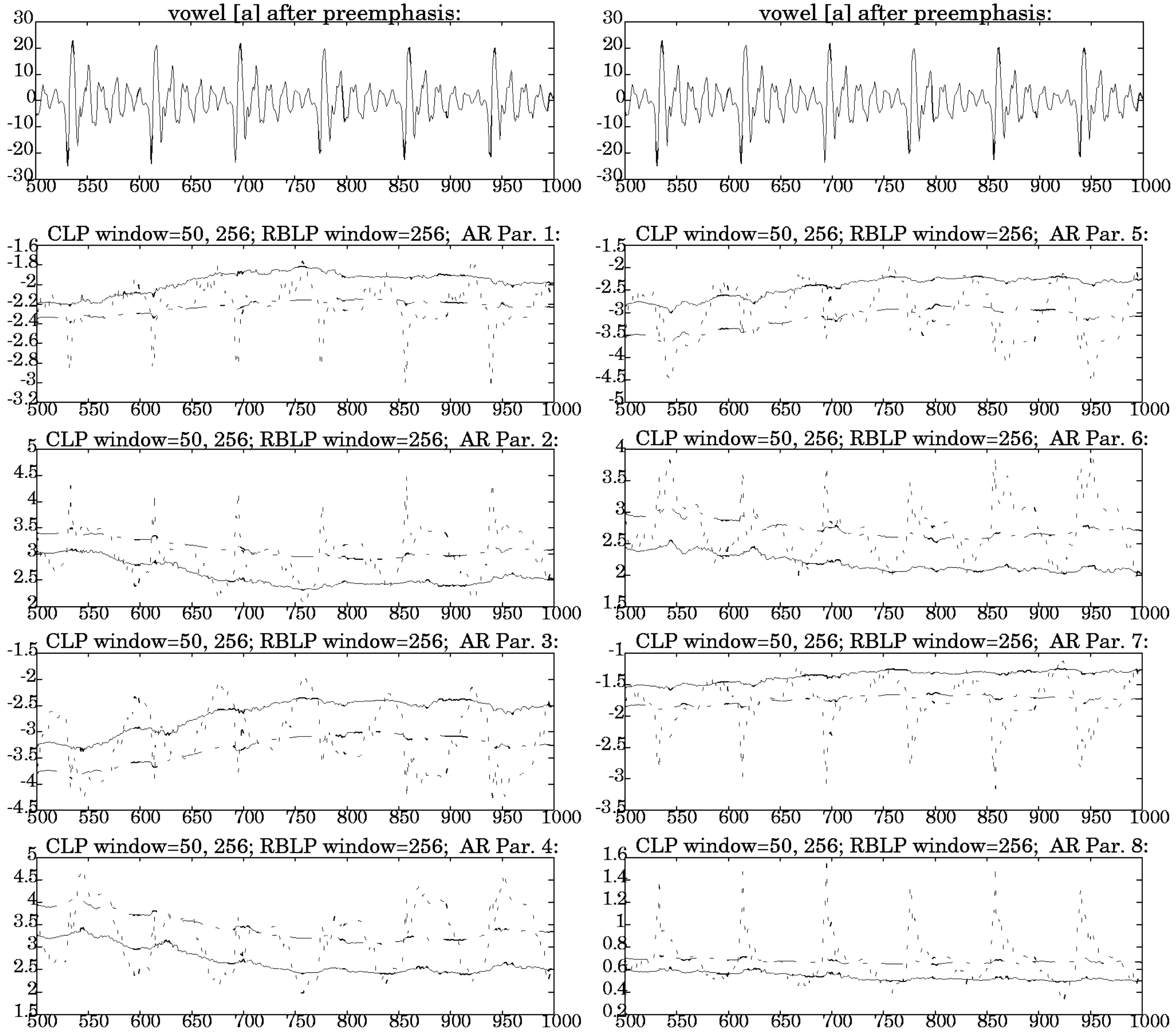

Figure 4.

trajectories (with the parts of the most accurate values), obtained using CLP method with sliding window shorter than the pitch period, while the dashdot and solid lines represent the parameter estimates for CLP and RBLP algorithm, respectively. These algorithms are based on the sliding window longer than the pitch period. The results presented indicate that RBLP algorithm tracks much better the reference trajectory, and it is more suitable for the natural human speech analysis than CLP estimates. Also it can be seen that both CLP and RBLP procedures give biased estimates, but the bias is smaller for RBLP algorithm. Similar results are also obtained for the vowels 'e', 'i', 'o' and 'u', but they are not presented owing to the space limitations.

4. CONCLUSIONS

In this paper we gave comparative analysis of standard CLP method and non-recursive RBLP procedure in estimating parameters of AR speech signal model. We have shown that RBLP algorithm produced less biased estimates of the LP coefficients than CLP method. Also, the RBLP algorithm produced a smaller variance estimates than CLP. The new estimator efficiency depends on degrees of freedom α. The experimental analysis shown that choice of α=3 leads to the most accurate and the most efficient estimate.

5. REFERENCES

[1] J.Markel. A.H.Gray, Jr., “Linear Prediction of Speech”, New York: Springer-Verlag. 1976

[2] J. Sanubari, K. Tokuda and M. Onoda, “Speech Analysis Based on AR Model Driven by t-Distribution Process,” *IEICE Trans.* Vol. E75-A, no. 9, pp. 1159-1169, September 1992.

[3] M.Veinović, B. Kovačević and M. Milosavljević, “Robust non-recursive AR speech analysis”, *Signal Processing*, Vol. 37, No. 2, May 1994, pp. 189-201