ON ROBUST SPEECH ANALYSIS BASED ON TIME-VARYING COMPLEX AR MODEL

Keiichi Funaki, Yoshikazu Miyanaga, Koji Tochinai

Department of Electronics & Information Engineering
Graduate School of Engineering, Hokkaido University
North-13 West-8, Kita-ku, Sapporo, 060-8628, Japan
funaki@media.eng.hokudai.ac.jp

ABSTRACT

We have already developed time-varying complex AR (TV-CAR) parameter estimation based on minimizing mean square error (MMSE) for analytic speech signal. Although the MMSE approach is commonly and successfully applied in various parameter estimation such as conventional LPC, it is well-known that an MMSE method easily suffers from biased and inaccurate spectrum estimation due to non-Gaussian nature of glottal excitation for voiced speech in the context of speech analysis. This paper offers robust parameter estimation algorithm for the TV-CAR model by applying Huber's robust M-estimation approach and two kinds of robust algorithms are derived: Newton-type algorithm and weighted least squares (WLS) algorithm. The preliminary experiments with synthetic signal generated by glottal source model excitation and natural speech uttered by female speaker demonstrate that the time-varying complex AR method is sufficiently robust against non-Gaussian nature of glottal source excitation owing to the improved resolution in the frequency domain.

1. INTRODUCTION

LPC methods[1][2] have been successfully utilized in a broad range of speech processing. The LPC methods, however, can not extract time-varying features from speech signal since observed speech signal is assumed as stationary within the analysis interval. On the other hand, several complex LPC methods for an analytic signal have already been proposed[3][4]. Analytic signal is a complex-valued signal whose real part is an observed signal and whose imaginary one is a Hilbert transformation of the observed signal. Since analytic signals provide the spectrum only in the positive frequency domain $(0, \pi/2)$, analytic signals can be decimated by a factor two. Consequently, these methods applying for analytic signal take some advantages over conventional real-valued LPC methods, i.e., more accurate spectral estimation, smaller computational amount, smaller errors in terms of computation with finite precision as well as quantization of the coefficients, and so on. We have already proposed a non-recursive complex speech analysis based on minimizing mean square error (MMSE) for analytic signal by introducing a time-varying complex AR (TV-CAR) model in which the parameters are represented by complex basis expansion[5]. In this method the complex AR coefficients can be efficiently estimated by solving linear equation by means of an extended version of LDU decomposition. The method can extract time-varying features from speech signal with non-recursive processing based on MMSE approach. The MMSE is optimal providing that the underlying distribution is represented by Gaussian. However it is well known that the outliers make it difficult to estimate accurate speech spectrum due to the non-Gaussian nature of glottal source excitation for voiced speech, especially in high-pitched speech. In order to realize robust estimation, Huber's robust M-estimation has been applied to LPC method[6][7]. In the robust estimation, the non-Gaussian nature of glottal excitation is assumed to be mixture distribution in which large portion of the excitations are from a normal distribution with a very small variance and a small portion of excitations are from an unknown distribution with a much bigger variance[6]. This distribution is often called heavy-tailed non-Gaussian. In this paper, we present the robust non-recursive speech analysis method based on the TV-CAR model for analytic speech signal by introducing Huber's robust M-estimation.

This paper is organized as follows. In section 2, the time-varying complex AR (TV-CAR) model is explained briefly. In section 3, robust M-estimation algorithm for the TV-CAR model is then derived. In the section, two robust M-estimation algorithms are derived: newton type algorithm and weighted least squares (WLS) algorithm. In section 4, experiments with synthetic signal driven by glottal source model excitation and natural speech uttered by female speaker are demonstrated.

2. TV-CAR MODEL

Target signal of the time-varying complex AR (TV-CAR) method is an analytic signal [8] that is complex-valued signal defined by

$$y^c(t) = \frac{y(2t) + jy_H(2t)}{\sqrt{2}} \tag{1}$$

where $y^c(t)$, $y(t)$, and $y_H(t)$ denote an analytic signal at time t, an observed signal at time t, and a Hilbert transformed signal for the observed signal $y(t)$, respectively. Since analytic signals hold the spectra only over the range $(0, \pi/2)$, analytic signals can be decimated by a factor

two. The term of $1/\sqrt{2}$ is multiplied in order to adjust the power of an analytic signal.

The introduced TV-CAR model[5] is defined as follows.

$$a_i^c(t) = \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) \quad (2)$$

$$H^c(z^{-1},t) = \frac{1}{1+\sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) z^{-i}} \quad (3)$$

where $H^c(z^{-1},t)$, $a_i^c(t)$, I, T, and $f_l^c(t)$ are taken to be a transfer function of the model, i-th complex AR coefficient at time t, AR order, finite order of complex basis expansion, and a complex-valued basis function, respectively. In the TV-CAR model, the complex AR coefficient is expressed with the finite number of complex basis function such as complex Fourier basis $exp(-j2\pi lt/T)$ or first order polynomial ($f_0^c(t)=1, f_1^c(t)=t$), or so on. In [5], we have derived the MMSE solution for the TV-CAR model, which is complex-valued LDU decomposition. Note that superscript c denotes complex value in this paper.

3. ROBUST ALGORITHMS

Huber's robust M-estimation[6][7] is applied to the previously proposed TV-CAR method in order to realize robust estimation. Huber's robust M-estimation is defined as the minimization of the sum of appropriately weighted prediction errors. The weight is a function of the prediction errors and the weight function is selected so as to down-weight the outliers appropriately.

$$E^c = \sum_{t=I}^{T-1} \rho\left[\frac{e_g^c(t)}{w}\right] \quad (4)$$

$$e_g^c(t) = y^c(t) + \sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) \quad (5)$$

Eq.(5) is the prediction error at time t for the feature vector $g_{i,l}^c$. In Eq.(4), $\rho[x]$ is called robust score function that cuts off the outliers of the non-Gaussian signal and w is scale factor that makes the criterion scale-invariant. The following Huber's score function is commonly adopted as robust score function.

$$\rho[x] = \begin{cases} C|x| - C^2/2 & (|x| \geq C) \\ x^2/2 & (|x| < C) \end{cases} \quad (6)$$

If $\rho[x]$ is $x^2/2$, this method is exactly equal to the MMSE-based TV-CAR method[5]. By taking the derivative of the weighted criterion Eq.(4), we can derive the following non-linear equation which requires iterative methods to solve.

$$\sum_{t=I}^{T-1} \psi\left[\frac{e_g^c(t)}{w}\right] \frac{f_n^c(t)^* y^c(t-k)^*}{w} = 0 \quad (7)$$

where $\psi[x]$ is the derivative of $\rho[x]$.

There are two approaches to solve Eq.(7), viz. newton-type algorithm and weighted least squares (WLS) algorithm.

3.1 Newton-type algorithm

$\psi\left[\frac{e_g^c(t)}{w}\right]$ in Eq.(7) can be approximated by first order Taylor series expansion.

$$\begin{aligned} \psi\left[\frac{e_g^c(t)}{w}\right] &\simeq \psi\left[\frac{\tilde{e}_g^c(t)}{w}\right] + \psi'\left[\frac{\tilde{e}_g^c(t)}{w}\right] \frac{e_g^c(t) - \tilde{e}_g^c(t)}{w} \\ &= \psi\left[\frac{\tilde{e}_g^c(t)}{w}\right] + \psi'\left[\frac{\tilde{e}_g^c(t)}{w}\right] \\ &\times \sum_{i=1}^{I}\sum_{l=0}^{L-1} (g_{i,l}^c - \tilde{g}_{i,l}^c) f_l^c(t) y^c(t-i)/w \end{aligned} \quad (8)$$

In Eq.(8), $\psi'[x]$ denotes the derivative of $\psi[x]$ that is called influence function, and $\tilde{g}$ denotes a preliminary estimation of $g_{i,l}^c$.

By substituting Eq.(8) into Eq.(7), we can obtain the following equation.

$$\begin{aligned} -\varphi(k,n) &= \sum_{i=1}^{I}\sum_{l=0}^{L-1} (g_{i,l}^c - \tilde{g}_{i,l}^c)\Psi(k,n,i,l) \\ \varphi(k,n) &= \sum_{t=I}^{T-1} \psi\left[\frac{\tilde{e}_g^c(t)}{w}\right] y^c(t-k)^* f_n^c(t)^* w \\ \Psi(k,n,i,l) &= \sum_{t=I}^{T-1} \psi'\left[\frac{\tilde{e}_g^c(t)}{w}\right] \\ &\times f_l^c(t) f_n^c(t)^* y^c(t-i) y^c(t-k)^* \\ &(1 \leq i,k \leq I, 0 \leq l,n < L) \end{aligned} \quad (9)$$

The equation is solved iteratively up to the enough convergence.

3.2 WLS algorithm

In Eq.(7), the following weighted function $W[x]$ is adopted.

$$W[x] = \frac{\psi[x]}{x} \quad (10)$$

By substituting Eq.(10) into Eq.(7) with approximation, we can obtain the following equation.

$$\sum_{t=I}^{T-1} W\left[\frac{e_g^c(t)}{w}\right]\left(\frac{e_g^c(t)}{w}\right) \frac{f_n^c(t)^* y^c(t-k)^*}{w} = 0 \quad (11)$$

$$\begin{aligned} -\varphi(k,n) &= \sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c \Psi(k,n,i,l) \\ \varphi(k,n) &= \sum_{t=I}^{T-1} W\left[\frac{e_g^c(t)}{w}\right] y^c(t) y^c(t-k)^* f_n^c(t)^* \\ \Psi(k,n,i,l) &= \sum_{t=I}^{T-1} W\left[\frac{e_g^c(t)}{w}\right] \\ &\times y^c(t-i) y^c(t-k)^* f_l^c(t) f_n^c(t)^* \\ &(1 \leq i,k \leq I, 0 \leq l,n < L) \end{aligned} \quad (12)$$

Eq.(11) can be solved with iteration by means of complex-valued LDU decomposition since the $\Psi(k,n,i,l)$ is an Hermit matrix.

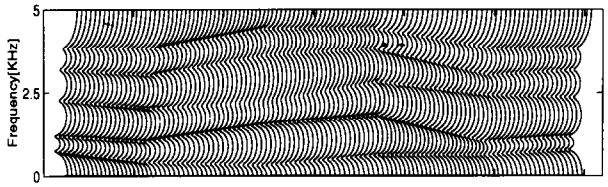
In WLS algorithm, more effective robust score function can be introduced, for example, Turkey's biweight function as follows.

$$\psi[x] = \begin{cases} x[1-(x/C)^2]^2 & (|x| \le C) \\ 0 & (|x| > C) \end{cases} \quad (13)$$

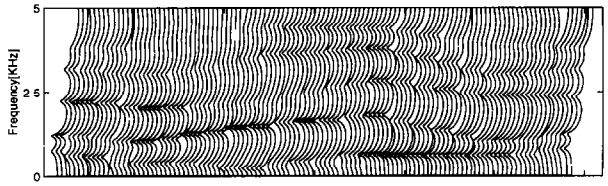
4. EXPERIMENTS

The experiments with synthetic signal driven by glottal source excitation and high-pitched natural speech were conducted. The testing synthetic signal is synthesized with time-varying ten order AR process /aiueoa/ by driving a glottal source excitation. The glottal excitation is generated by Rosenberg-Klatt model (RK-model)[9] with the parameters $(AV, OQ, TL) = (200, 0.75, 10)$ and pitch period $T0 = 5[msec]$. The reference spectrum of the AR process is shown in Fig.1(a1). The AR parameters of the synthetic signal are linearly interpolated with the corresponding formant frequency and bandwidth between typical all-pole spectra located at every $50[msec]$ interval to generate a time-varying spectrum. Sampling rate of the synthetic signal is supposed to be $10[kHz]$. The testing natural speech /ge/ is drawn in Fig.2(a2). The signal is $10[KHz]$ sampled speech that is converted from $20[KHz]$ sampled ATR database data and its speaker is FKN. Table 1 shows analysis conditions. In Table 1, T and S denote analysis width and shift length($[msec]$). In Table 1, (b)-(g) means as follows. (b) is most popular autocorrelation LPC method. (c) is time-varying covariance LPC method that is real-valued version of the MMSE method[5]. (d) is robust estimation algorithm of (c). (e) is complex covariance LPC method. (f) is MMSE TV-CAR method[5]. (g) is proposed robust TV-CAR method. In the robust methods (d) and (g), the robust estimation is realized by **3.2** WLS algorithm with Tukey's biweight function, $C = 1.5$, and iteration number is 2. In the time-varying methods (c),(d),(f),(g), first order polynomial is adopted as basis function, i.e. $f_l^c(t) = t^l/l$. Note that pre-emphasis operation is not introduced in any methods. Analysis order is 14 for the real-valued methods and 7 for the complex-valued methods. Moreover, (20,20) IIR filter [10] is adopted to realize Hilbert transform.

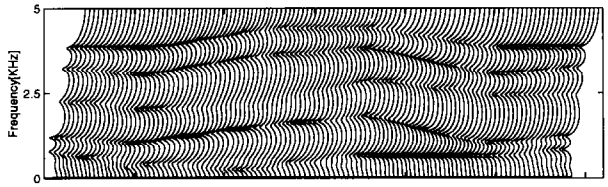
Fig.1 and Fig.2 show the estimated spectra with synthetic signal and natural speech, respectively. In both figures, spectrum is drawn at every $2[msec]$. (b) can only estimate one spectrum for one analysis frame, thus, the same spectrum is repeatedly drawn within the same analysis frame. Fig.1 and 2 demonstrate that robust estimation is not so effective for complex-valued method although robust real-valued method can estimate less biased and less valiance spectrum than non-robust one. The reason is that the resolution in the frequency domain on complex-valued method is improved twice than that on real-valued one owing to the decimation with factor two. Furthermore, a basis function constrains the parameters to vary in time in the TV-CAR method. The constraint leads to less variance spectrum estimation. Consequently, the TV-CAR method is enough robust against the non-Gaussian nature of glottal source excitation.



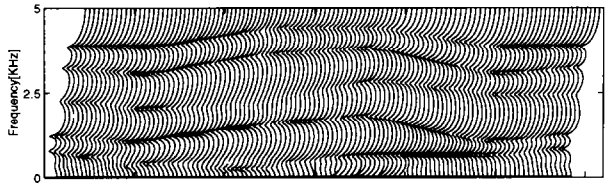
(a1) Reference spectra /aiueoa/



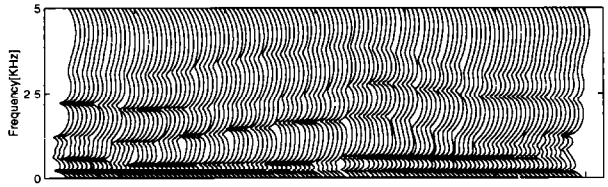
(b) LPC($T = 20, S = 10$,Hamming window)



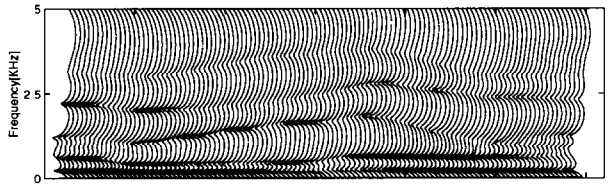
(c) Time-varying covariance($T = 20, S = 10$)



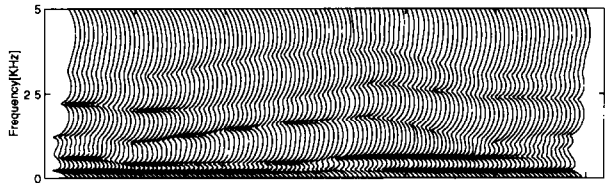
(d) Robust Time-varying($L = 2, T = 20, S = 10$)



(e) Complex covariance($L = 1, T = 20, S = 10$)



(f) TV-CAR($L = 2, T = 20, S = 10$)



(g) Robust TV-CAR($L = 2, T = 20, S = 10$)

Fig.1 Experimental results with high-pitch synthetic speech /aiueoa/ generated by RK-model excitation

Table 1 Analysis conditions

	Method	L	T	S
(b)	Auto-correlation LPC[1]	-	20	10
(c)	Time-varying covariance	2	20	10
(d)	Robust time-varying covariance	2	20	10
(e)	Complex covariance	1	20	10
(f)	TV-CAR [5]	2	20	10
(g)	Robust TV-CAR	2	20	10

5. CONCLUSIONS

Robust M-estimation has been applied to the time-varying complex AR (TV-CAR) method, which can take into account the non-Gaussian nature of glottal source excitation. The preliminary experimental results with synthetic signal and natural speech demonstrate that the TV-CAR method is sufficiently robust against the non-Gaussian nature of glottal excitation since the resolution in the frequency domain is improved twice due to the decimation of analytic signals with a factor two and AR parameters are constrained to vary in time by basis function in the TV-CAR method. Evaluating the robust TV-CAR method in noisy environment is future study.

6. ACKNOWLEDGEMENT

The authors would like to thank to Dr. M.Hiroshige of Hokkaido University for his support of this work.

7. REFERENCES

[1] F.Itakura and S.Saito, "A statical method for estimation of speech spectral density and formant frequency," IEICE Trans., Vol.53-A, pp.35-42, 1970. (in Japanese)

[2] B.S.Atal and S.L.Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J.Acoust. Soc. Am., Vol.50, pp.637-644, 1971.

[3] S.M.Kay, "Maximum entropy spectral estimation using the analytic signal," IEEE Trans. ASSP-26, pp.467-469, 1980.

[4] T.Shimamura et.al., "Complex linear prediction method based on positive frequency domain," IEICE Trans., Vol.J72-A, pp.1755-1763, 1989. (in Japanese)

[5] K.Funaki et.al.,"On a time-varying complex speech analysis,"EURASIP Proc. Eusipco-98, Sep. 1998.

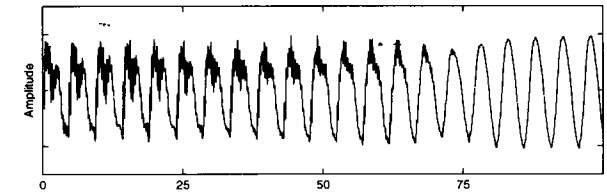
[6] C-H Lee, "On robust linear prediction of speech," IEEE Trans. ASSP-36, 1988.

[7] M.D.J.Veinovic et.al.,"Robust non-recursive AR speech analysis," Signal Processing Vol. 37 pp.189-201. 1994.

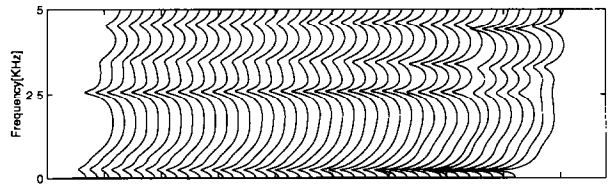
[8] A.V.Oppenheim and R.W.Schafer, "Digital signal processing," Prentice Hall, 1975.

[9] D.Klatt and L.Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J.Acoust. Soc. Am., Vol.87, pp.820-857, 1990.

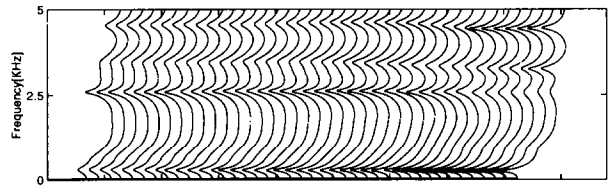
[10] M.Ikehara et.al., "Design of IIR Hilbert transformers using Remez algorithm," IEICE Trans., Vol.J74-A, pp.414-420, 1991.(in Japanese)



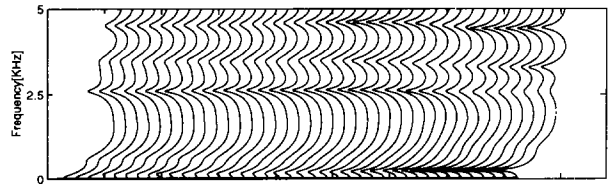
(a2) Natural Speech /ge/



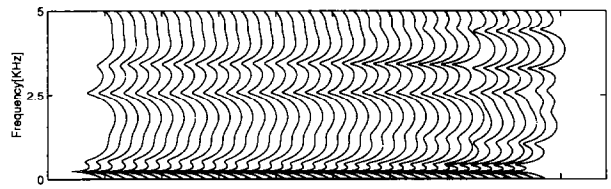
(b) LPC($T = 20, S = 10$,Hamming window)



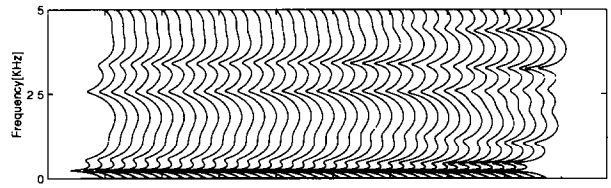
(c) Time-varying covariance($T = 20, S = 10$)



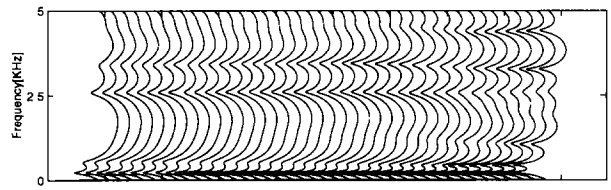
(d) Robust Time-varying($L = 2, T = 20, S = 10$)



(e) Complex covariance($L = 1, T = 20, S = 10$)



(f) TV-CAR($L = 2. T = 20, S = 10$)



(g) Robust TV-CAR($L = 2, T = 20, S = 10$)

Fig.2 Experimental results with natural speech /ge/ uttered by female speaker