NOISE ROBUST FORMANT FREQUENCY ESTIMATION BASED ON COMPLEX AUTOCORRELATION FUNCTION

Ousmane Diankha and Tetsuya Shimamura
Department of Information and Computer Sciences
Saitama University
255 Shimo-Okubo, Saitama 338-8570, Japan
Tel: +81-48-858-3496, Fax: +81-48-858-3716
e-mail: diankha@sie.ics.saitama-u.ac.jp, shima@sie.ics.saitama-u.ac.jp

Abstract: This paper proposes an improved method for formant frequencies estimation based on the complex autocorrelation function of the speech signal. Instead of using the incoming signal as an input for the LPC analysis, the analytic signal of the autocorrelation function of the speech signal is computed and itself used as an input for the LPC analysis. Due to the properties of the analytic signal, which occupies half of the bandwidth of the original signal, the required model order for the LPC analysis is halved. The accuracy of the proposed method in noisy environments is examined on five natural vowels. The effectiveness of the proposed method is shown by the estimated spectral shapes and the estimation errors of the formant frequencies.

1. Introduction

Formant frequency estimation plays an important role in many applications of speech signal processing: speech recognition, speaker identification and related areas. Formants frequencies represent the most immediate source of articulatory information and are critical in speech perception.

The most common approach for formant frequency estimation is the Linear Predictive Coding (LPC) method [1]-[3], which can extract the formant frequencies effectively by finding the roots of the linear predictor polynomial, or by picking the peaks of the LPC spectrum. However, in noisy environments, the LPC method is affected by noise and less effective. To solve this problem, Duncan et al. [4] proposed a pole focusing method. The method unfortunately requires complicated calculations.

Recently, an effective and efficient formant estimation method in noisy environments was proposed by Zhao et al. [5]. It was shown in [5] that the noisy part of the autocorrelation function (ACF) of the corrupted speech signal is easily located. By utilizing this property, a method using the ACF of the noisy speech signal as the input for the LPC method was derived and demonstrated, in which the method provided a significant improvement relative to the standard LPC method at a moderate signal-to-noise ratio (SNR). At a (very) low SNR, however, the performance of the ACF based method is competitive with that of the standard LPC method.

To overcome the drawback, we propose an improved method in this paper. It consists of using the analytic signal that has important properties for speech signal processing. The key point of this approach is to diminish the noisy part included in the ACF of the corrupted speech signal when generating the analytic signal. And the analytic signal filters out the negative components of the frequency axis, avoiding the interaction between positive and negative frequency components. This leads to an accurate formant frequency estimate.

The rest of the paper is organized as follows. The proposed method is described in Section 2, pointing out the merit of the analytic signal transformer. Experiments conducted are described in Section 3. Section 4 is devoted to drawing conclusions.

2. Proposed Method

A block diagram of the proposed method is depicted in Fig.1. The process is divided into three main steps: the pre-treatment that consists of the pre-emphasis and autocorrelation function computation; the second phase, analytic signal generation, is the core of our method. Finally, the complex LPC analysis [6] is applied to detect formant frequencies.

Let $s_R(n)$ be the incoming speech signal sampled. The input signal is first pre-processed to reduce the slope of the speech spectrum. This is implemented by a pre-emphasis filter having the following form:

$$P(z) = 1 - \mu z^{-1} \qquad (1)$$

where μ corresponds to the pre-emphasis factor.

A Hamming windowing function is used on the pre-emphasized speech signal $\tilde{s}_R(n)$ so as to reduce the edge effects at the beginning and the end of the frame. And, the ACF of the windowed signal $\bar{\tilde{s}}_R(n)$ is calculated as

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} \bar{\tilde{s}}_R(n) \bar{\tilde{s}}_R\ (n+k) \quad k{=}0,1,..p \qquad (2)$$

where N is the data length in the frame and k the lag value. $R(k)$ satisfies the following properties:

$a)$ $R(k)$ is composed of the same frequency components as $s_R(n)$.

$b)$ The amplitude of each frequency component of $R(k)$ is proportional to the square of that of $s_R(n)$.

$c)$ If $s_R(n)$ is white noise, then the energy of $R(k)$ is concentrated on $k = 0$.

Then, $R(k)$ is passed through the analytic signal transformer (AST) and a complex ACF (CACF) is obtained The CACF results in an analytic (complex) signal whose components are Hilbert transforms of each other for the real and imaginary parts [7].

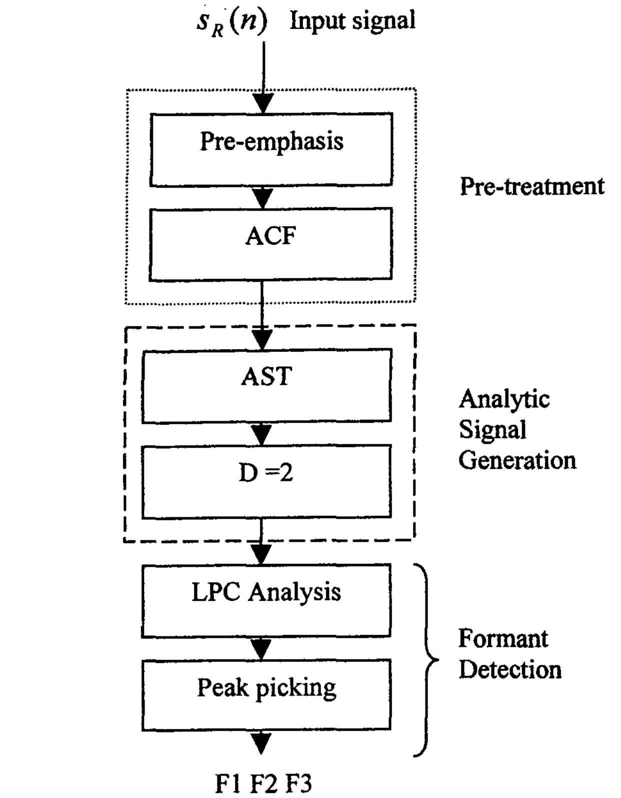


Fig.1: Block diagram of the proposed method

Let $R_a(k)$ be the CACF, the analytic signal corresponding to $R(k)$; it is expressed as follows:

$$R_a(k) = R(k) + jH(R(k)) \tag{3}$$

where $H(R(k))$ denotes the Hilbert transform of $R(k)$. $R(k)$ and $H(R(k))$ are related as follows:

$$H(R(k)) = \frac{1}{\pi}\sum_{\substack{m=-\infty \\ m\neq k}}^{\infty} R(k-m)\frac{\sin^2[\pi(m/2)]}{m} \tag{4}$$

and

$$R(k) = -\frac{1}{\pi}\sum_{\substack{m=-\infty \\ m\neq k}}^{\infty} H(R(k-m))\frac{\sin^2[\pi(m/2)]}{m} \tag{5}$$

In the step of obtaining the CACF, the AST can be implemented by using a complex FIR filter. Fig.1 shows a block diagram of the AST used in this paper.

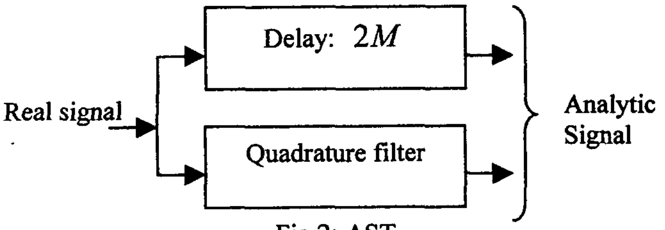


Fig.2: AST

It contains two parts: the upper part producing a delay of $2M$ in the case of the filter length $l = 4M + 1$, and the lower part producing the imaginary part by an FIR quadrature filter. The output sequences of the AST form the real and imaginary parts of $R_a(k)$. The filter structure of AST in Fig.2 essentially produces an output delayed by $2M$; this works effectively to diminish the noisy part included in the input ACF, because according to the property c) of the ACF described above there is the noise component only in $R(0)$ if the additive noise is white (even if the noise is colored, it may be concentrated in near $R(0)$).

As $R_a(k)$ occupies half of the bandwidth of $R(k)$, $R_a(k)$ can be down-sampled by a factor of 2. The complex LPC method [6] is applied to the down-sampled $R_a(k)$. From the resulting complex LPC spectrum, the first three formants are searched based on the peak-picking algorithm. The candidates of each frequency F1, F2 and F3 are then validated according to a predefined interval for each of these frequencies [3].

According to the properties a) and b) of the ACF described above, the formant frequencies of the input noisy speech can be obtained from $R_a(k)$, though $R_a(k)$ is the down-sampled analytic signal of $R(k)$. In addition, the SNR of $R_a(k)$ is increased relative to that of $R(k)$ due to the effect of diminishing the noise in AST. Thus, it is expected that the proposed method more accurately estimates the formant frequencies of noisy speech.

3. Simulation Results and Discussions

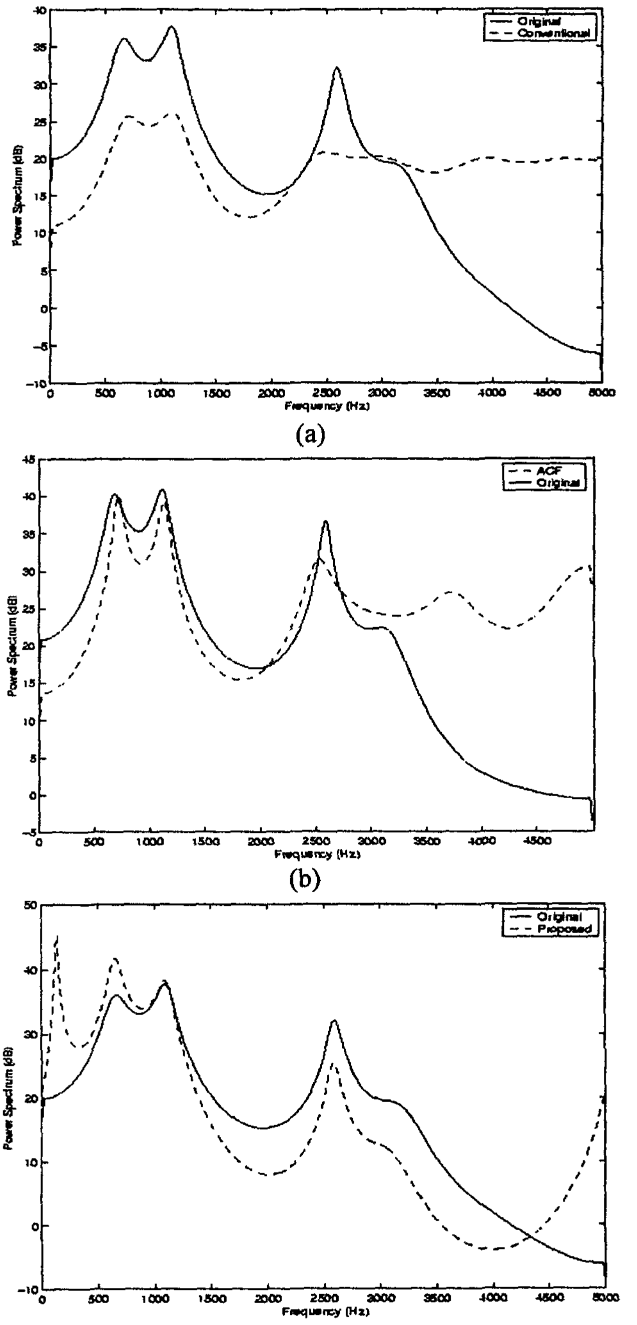
To evaluate the effectiveness of the proposed method, five natural vowels "*e, a, u, o, i*" uttered by a male speaker are analyzed.

First, preliminary computations of the conventional LPC method [1] for 10 frames in noiseless environments are made to determine the three first formants (F1, F2, F3) of each utterance. The average of 10 frames is used as the reference for each formant frequency. Then, each utterance is corrupted by a Gaussian white noise with SNRs of 5 and 10 dB. The simulation parameters are listed in Table1.

Table1: Simulation parameters

Sampling frequency	10kHz
Pre-emphasis factor	0.975
Window length	Hamming, 51.2 ms
LPC order	12
CLPC order	6
SNR	5dB, 10dB

An example of the spectral envelopes obtained by the conventional LPC method [1], ACF-LPC method [5] and proposed method on the vowel /*a*/ at a SNR of 10 dB is shown in Fig.3.



(a)

(b)

(c)

Fig.3: Comparison of the LPC spectra estimated on vowel /a/. (a) conventional LPC method (b) ACF-LPC method (c) proposed method

Fig.3 (a) compares the original LPC spectrum of a noise free speech signal (solid line) with the spectrum obtained by the conventional LPC method on noisy speech (dashed line). A performance degradation is obviously observed. Fig.3 (b) compares the original LPC spectrum of the noise free speech signal (solid line) with the spectrum obtained by the ACF-LPC method on noisy speech (dashed line). Here, a deviation should be noted in the three formants and particularly in the third formant. Fig.3 (c) compares the original LPC spectrum of the noise free speech signal (solid line) with the spectrum obtained by the proposed method on noisy speech (dashed line). In Fig.3 (c), the first peak has occurred at a frequency range less than the minimum value of F1. However, this is not considered as a formant in the proposed method, because a frequency range is fixed for each formant (F1, F2, and F3) in the proposed method. By inspecting the spectral envelopes in Fig.3, we see that the proposed method has the potential to extract formant frequencies from a noisy speech signal more accurately than the conventional LPC and ACF-LPC methods.

As numerical results for each formant frequency, the estimation error Δ is evaluated where Δ is expressed as follows:

$$\Delta = \frac{|F_{iE} - F_{iT}|}{F_{iT}} \times 100 \qquad (6)$$

where F_{iE} are the estimated values of the i-th formant and F_{iT} the references.

Table2: Estimation errors by the conventional LPC method at 10dB

Vowels	Frequencies		
	F1	F2	F3
e	5.38	1.05	19.39
a	1.79	0.96	5.07
u	14.18	15.27	2.93
o	23.41	54.17	15.50
i	65.50	37.30	17.88
average	22.05	21.75	12.15

Table3: Estimation errors by the ACF-LPC method at 10dB

Vowels	Frequencies		
	F1	F2	F3
e	3.32	2.76	3.00
a	1.78	2.51	1.01
u	4.47	14.21	2.54
o	3.01	14.11	3.05
i	89.60	42.91	29.31
average	20.43	15.30	7.78

Table4: Estimation errors by the proposed method at 10dB

Vowels	Frequencies		
	F1	F2	F3
e	1.67	2.24	7.78
a	0.69	2.09	1.20
u	0.75	10.29	0.82
o	15.37	70.34	11.21
i	21.34	17.90	7.01
average	7.96	20.57	5.60

Tables 2, 3 and 4 show the results obtained by the conventional LPC method, the ACF-LPC method, and the proposed method for the case of SNR=10 dB, respectively.

The total performance of the three methods is evaluated for each frequency F1, F2 and F3 by taking the average error of the five vowels in each of these frequencies. It can be seen that the proposed method outperforms the conventional LPC and ACF-LPC methods, although the proposed method fails particularly for the second formant of the vowel /o/. To confirm this, a more severe SNR of 5dB is further investigated. In this situation, it is seen that the performance of the conventional LPC method is drastically decreased as shown in Table 5. The performance of the ACF-LPC method also decreases as shown in Tables 6. The results of the proposed method in Table 7 give the same phenomena noted for the vowel /o/ with SNR=10 dB. This suggests that the failure of the proposed method for the vowel /o/ may not be attributed to the performance of the method but to the signal behavior itself. It is, however, observed that the proposed method provides more accurate estimates again.

Table5: Estimation errors by the conventional LPC method at 5dB

Vowels	Frequencies		
	F1	F2	F3
e	1.89	2.75	4.00
a	11.11	115.79	26.59
u	3.45	23.33	27.35
o	20.93	2.50	6.51
i	89.56	55.43	37.08
average	25.38	39.46	20.30

Table6: Estimation errors by the ACF-LPC method at 5dB

Vowels	Frequencies		
	F1	F2	F3
e	2.59	2.28	16.55
a	5.63	1.99	4.19
u	26.31	1.84	0.08
o	3.42	11.97	15.26
i	88.80	46.83	32.00
average	25.35	12.98	13.61

Table7: Estimation errors by the proposed method at 5dB

Vowels	Frequencies		
	F1	F2	F3
e	1.89	2.29	0.36
a	1.39	1.75	9.36
u	3.45	0.83	19.59
o	9.30	18.75	3.83
i	89.56	27.27	12.53
average	21.11	10.17	9.13

4. Conclusion

In this paper, an improved method for formant frequencies estimation using the complex autocorrelation function has been presented. And, it has been demonstrated that the proposed method outperforms both the conventional LPC and ACF-LPC methods in noisy environments.

References

[1] J. D. Markel, *"Digital inverse filtering - A new tool for formant trajectory estimation"* IEEE Trans. Audio and Electroacoustics, Vol. AU-20, no 2, pp.129-137. 1972.

[2] S. Furui, *" Digital Speech Processing, Synthesis, and Recognition"* Marcel Dekker, 1989.

[3] L. Rabiner and B-H. Juang, *"Fundamentals of Speech Recognition*", Englewood Cliffs: Prentice Hall, 1993.

[4] G. Duncan and M.A. Jack, *"Formant estimation algorithm based on pole focusing offering improved noise tolerance and feature resolution",* IEE Proceedings, Vol.35, Pt.F, No.1, pp.18-32, 1988.

[5] Q. Zhao and T. Shimamura, *"A robust algorithm for formant frequency extraction of noisy speech",* Proceedings of IEEE International Symposium on Circuits and Systems, pp.534-537, 1998.

[6] T. Shimamura and S. Takahashi: *"Complex linear prediction method based on positive frequency domain"* Electronics and Communications in Japan, Part 3,vol. 73 no 9, pp. 1990; translated from Denshi Joho Tsushin Gakai Ronbunshi, Vol. 72A, no. 11, pp.1755-1763, 1989.

[7] M. Bellanger: *"Digital Processing of Signals: Theory and Practice",* John Wiley & Sons, 1989.