Efficient Tracking of Speech Formant Using Closed Phase WRLS-VFF-VT Algorithm

*Kyo-Sik Lee, **Kyu-Sik Park

Abstract

In this paper, we present an adaptive formant tracking algorithm for speech using closed phase WRLS-VFF-VT method. The pitch synchronous closed phase methods is known to give more accurate estimates of the vocal tract parameters than the pitch asynchronous method. However the use of a pitch-synchronous closed phase analysis method has been limited due to difficulties associated with the task of accurately isolating the closed phase region in successive periods of speech. Therefore we have implemented the pitch synchronous closed phase WRLS-VFF-VT algorithm for speech analysis, especially for formant tracking. The proposed algorithm with the variable threshold(VT) can provide a superior performance in the boundary of phone and voiced/unvoiced sound. The proposed method is experimentally compared with the other method such as two channel CPC method by using synthetic waveform and real speech data. From the experimental results, we found that the block data processing techniques, such as the two-channel CPC, gave reasonable estimates of the formant/antiformant. However, the data windows used by these methods included the effects of the periodic excitation pulses, which affected the accuracy of the estimated formants. On the other hand the proposed WRLS-VFF-VT method, which eliminated the influence of the pulse excitation by using an input estimation as part of the algorithm, gave very accurate formant/bandwidth estimates and good spectral matching.

I. Introduction

From our experience with a simplified cascade-parallel model of the vocal tract it appeared that the extraction of synthesizer control parameters from the LPC analysis of natural speech was quite complicated. Only in the vowels could the LPC formant be mapped onto the parameters of the cascade branch. In all other sounds, the approximation of the natural speech spectrum by selecting appropriate values of the parameters of the cascade and parallel branches appeared to be an art rather than a science. It was felt that a pole-zero model of the short-time speech spectrum would provide a much better fit to the acoustic behavior of the actual speech production apparatus, thus enabling us to interpret poles resulting from the analysis as genuine formant and to model anti-resonances in all speech sounds where they may occur. Using a cascade pole-zero model for synthesis would make large amounts of articulatory based phonetic knowledge available for guiding rule development.

For a nonstationary speech signal, accurate tracking speech parameters like vocal tract resonance frequencies(formants) and their bandwidths are essential for the development of speech recognition and speech synthesis systems. Using the frame-based linear predictive coding(LPC) analysis, the accuracy of formant tracking of a speech signal is affected by 1) the position of the analysis frame, 2) the length of the analysis windows, and 3) the time-varying characteristics of the speech signal. An adaptive filter approach, which tracks the time-varying parameters of the vocal tract and updates the parameters during the glottal closed phase interval, can reduce the formant estimation error.

Over the years there has been considerable interest in the time-varying modeling of speech signals[1][2]. The motivation behind this effort stems from the fact that many elements of speech such as stops, fricative onsets, and transitions between consonants and vowels, exhibit rapid changes that cannot be modeled satisfactorily by standard time-invariant techniques. In other words, an invariant model applied to relatively short segments of data is not appropriate in these situations. As a result, more meaningful feature sets for such events could be obtained by means of parametric models based on time-dependent difference equations which have been proposed in the recent literature[3]. However, most methods employ least squares techniques for parameter

* Dept. of Information and Telecommunication, Hansei University
** Dept. of Information and Telecommunication, Sangmyung University

Manuscript Received : April 14, 2000.

extraction and these do not perform very well in the presence of additive noise or when the magnitudes of some of the coefficients are very small, both of which are distinct possibilities when dealing with real speech data. Furthermore, these models fail to differentiate regions of glottal excitation from those where the glottis is closed.

In order to estimate the vocal tract parameters accurately, it is necessary to eliminate the influence of the pitch period from the spectra of speech signals. Input pulses are estimated from the speech waveform and this estimated input is used for the model input, we can expect to obtain correct formants without the influence of pitch, that is, the correct parameters of the speech production model. In ref[4], we proposed WRLS-VFF-VT method for estimating ARMA parameters, input pulse train, and input white noise at the same time so that formants and antiformants of speech can be correctly estimated. Using this WRLS-VFF-VT method, a variable forgetting factor is used to allow the estimation process to track the time-varying parameters accurately and even more quickly.

This paper is organized as follows. In section 2. we describes the basic WRLS-VFF-VT algorithm and implementation procedure for closed phase formant tracking. Section 3 present performance evaluation of the proposed algorithm compared with a previous two channel CPC method. Finally section 4 concludes this paper.

II. Adaptive Formant Tracking using Closed Phase WRLS-VFF Algorithm

Past use of a pitch-synchronous closed phase analysis method has been limited due to difficulties associated with the task of accurately isolating the closed phase region in successive periods of speech. Since the pitch synchronous closed phase methods is known to give more accurate estimates of the vocal tract parameters than the pitch asynchronous method, we have implemented the following pitch synchronous closed phase WRLS-VFF-VT algorithm for speech analysis. In this section, we will first describe how the adaptive WRLS-VFF algorithm works and we then show how this algorithm can be implemented for tracking speech formant efficiently.

2.1. Adaptive WRLS-VFF-VT Algorithm

Figure 1 shows the basic block diagram of the WRLS-VFF-VT algorithm for the ARMA parameter estimation.

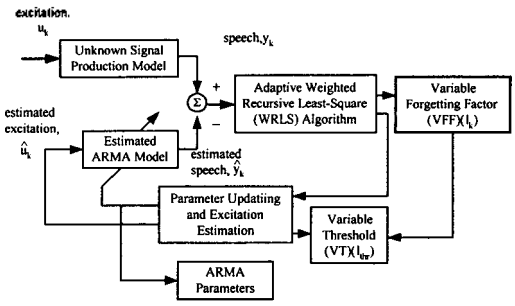

Figure 1. WRLS-VFF-VT algorithm for the ARMA parameter.

In the figure, we assume that the speech signal is generated by an ARMA model represented by the following :

$$y_k = -\sum_{i=1}^{p} a_i(k) y_{k-i} + \sum_{j=1}^{q} b_j(k) u_{k-j} + u_k \qquad (1)$$

Where y_k denotes the k-th sample of the speech signal, u_k is the input excitation to the ARMA model, (p,q) are the order of the poles and zeros, respectively, of the ARMA model, and $a_i(k)$ and $b_j(k)$ are the time-varying AR and MA parameters, respectively. Here, we assume that the values of p and q can be predetermined. Note that the measured speech signal, y_k, depends of the input, u_k. The excitation, u_k, is usually considered to be white Gaussian noise. We must estimate the input excitation, u_k, so that the ARMA parameters can be estimated accurately from y_k. An estimation method for u_k based on the variable forgetting factor of the WRLS algorithm will be given later. For the present we assume that an estimate for u_k is available.

Let us define ARMA parameter vector, θ_k, its estimate, $\hat{\theta}_k$, and a data vector, Φ_k, by the following equations :

$$\theta_k{}^t = [a_1(k), \ldots, a_p(k), b_1(k), \ldots, b_q(k)] \qquad (2)$$

$$\hat{\theta}_k{}^t = [\hat{a}_1(k), \ldots, \hat{a}_p(k), \hat{b}_1(k), \ldots, \hat{b}_q(k)],$$
$$\Phi_k{}^t = [-y_{k-1}, \ldots, -y_{k-p}, u_{k-1}, \ldots, u_{k-q}]$$

where the superscript t denotes transpose, and $\hat{a}_i$ and $\hat{b}_i$ are the estimated ARMA parameters, respectively. Using (2) the speech signal, y_k, and its estimate, $\hat{y}_k$, and the residual error of the ARMA process, r_k, may be expressed as

$$y_k = \Phi_k{}^t\theta_k + u_k,\ \hat{y}_k = \hat{\Phi}_k{}^t\hat{\theta}_k + \hat{u}_k,\ r_k = y_k - \Phi_k{}^t\hat{\theta}_k \tag{3}$$

The weighted sum of the squares of the residual error can be expressed recursively as[10]

$$V_k(\theta) = \lambda V_{k-1}(\theta) + \xi_k^2(1 - \phi_k' K_k) \tag{4}$$

A strategy for calculating λ_k may be defined by requiring $V_k(\theta)$ to be constant. In other words, the forgetting factor will compensate at each step k for the new error information in the latest measurement, thereby insuring that the estimation is always based on the same error information. Thus from (4) by setting $\lambda = \lambda_k$, we have

$$\lambda_k = 1 - \frac{\xi_k{}^2}{V_1(\theta)}[1 - \Phi_k{}^t K_k] \tag{5}$$

For some applications, if λ_k becomes small, we recommend a minimal λ_k be defined as

$$\lambda_{\min} = 1 - \frac{1}{N_a},\ \text{if } \lambda_k < \lambda_{\min},\ \textit{then } \lambda_k = \lambda_{\min} \tag{6}$$

where Na = p + q is the total number of the filter coefficients in the ARMA model.

We now note that the measured speech signal, y_k, depends of the input, u_k and we must estimate the input excitation, u_k so that the ARMA parameters can be estimated accurately from y_k. For the estimation of input excitation u_k based on the variable forgetting factor of the WRLS algorithm is described as following. The input excitation u_k to a speech production process can be either pulse trains for voiced sounds or white noise for fricatives. Therefore a general expression of the input sequence is given as follows

$$u_k = u_k{}^p + u_k{}^w \tag{7}$$

where $u_k{}^p$ represents the pulse input and $u_k{}^w$ is the white noise input.

We introduce a weighted least squares criterion(cost function) to estimate the ARMA parameters based on the estimation error e_k (different from the residual error r_k), from equation (3), we have

$$e_k = y_k - \hat{y}_k = y_k - \phi_k{}^t\hat{\theta}_k - \hat{u}_k \tag{8}$$

Thus the cost function $V_k(\theta)$ is expressed as

$$V_k(\theta) = \sum_{i=1}^{k} \lambda^{k-1}\ (y_1 - \phi_i{}^t\hat{\theta}_i - \hat{u}_i)^2 \tag{9}$$

From (7) we have $\sum \hat{u}_k = \sum \hat{u}_k{}^w + \sum \hat{u}_k{}^p$, and under the theorem of ergodicity, for large k, these sums are expectations, which indicates that when the input to the reference model is white, the prediction error produced by an optimal prediction (e.g., WRLS algorithm) is also white. Thus, for a zero mean white noise input, we have $E[\hat{u}_k{}^w] = E[u_k{}^w] = 0$. Then (9) can be rewritten as

$$V_k(\theta) = \sum_{i=1}^{k} \lambda^{k-1}\ (y_1 - \phi_i{}^t\hat{\theta}_i - \hat{u}_i{}^p)^2 \tag{10}$$

The ARMA parameter vector θ_k that minimizes the cost function $V_k(\theta)$ can be obtained by setting the parallel derivative of V with respect to θ equal to zero, which gives the recursive equation to update $\hat{\theta}_k$ as

$$\hat{\theta}_k = \hat{\theta}_{k-1} - K_k\ (y_k - \phi_k{}^t\hat{\theta}_{k-1} - \hat{u}_k{}^p) \tag{11}$$

The proposed input estimation method uses the FF as a reference to examine the input condition. This can reduce the system complexity since only one adaptive algorithm is used instead of two as in [8]. Moreover, the FF can be obtained from the adaptive process. Thus no extra calculations are required.

Consequently, the WRLS-VFF algorithm can be specified by a set of recursive least square equations obtained by minimizing the cost function with respect to the parameter vector θ_k. This implementation of the WRLS-VFF-VT algorithm with input estimation is given in Table I.

Table 1. Implementation of the WRLS-VFF-VT algorithm with input estimation.

Prediction error : $\xi_k = y_k - \hat{\Phi}_k{}^t\hat{\theta}_{k-1}$

Gain : $K_k = P_{k-1}\hat{\Phi}_k[\lambda_{k-1} + \hat{\Phi}_k{}^t P_{k-1}\hat{\Phi}_k]^{-1}$

Forgetting Factor : $\lambda_k = 1 - \xi_k{}^2(1 - \hat{\Phi}_k{}^t K_k)^2 / E_1$

Threshold : $E_k = 1/M \sum_{i=1}^{M-1} \lambda_{k-i}$

If $E_k < 0.9$, then $\lambda_{thr} = 0.99 * E_k$

If $E_k > 0.9$, then $\lambda_{thr} = 0.9 * E_k$

If $\lambda_{thr} > \lambda_{\min}$, then $\lambda_{thr} = \lambda_{\min}$

Input estimate : a) If $\lambda_k < \lambda_{thr}$, then the input is a pulse and

$$\hat{u}_k{}^w = 0$$

$$\hat{u}_k = \hat{u}_k{}^p = y_k - \Phi_k{}^t \hat{\theta}_{k-1}$$

b) If $\lambda_k > \lambda_{thr}$, then the input is white noise and

$$\hat{u}_k{}^p = 0$$

$$\hat{u}_k = \hat{u}_k{}^w = \xi_k(1 - \Phi_k{}^t K_k)$$

Real prediction error : $e_k = (y_k - \Phi_k{}^t \hat{\theta}_{k-1} - \hat{u}_k{}^p)$

Parameter : $\hat{\theta}_k = \hat{\theta}_{k-1} + K_k e_k$

Covariance matrix : $P_k = \lambda_k{}^{-1}[P_{k-1} - K_k \Phi_k{}^t P_{k-1}]$

2.2. Closed Phase WRLS-VFF-VT algorithm

The closed phase WRLS-VFF-VT algorithm for speech analysis extracts the vocal tract parameters only from the glottal closed interval. The selection of the AR or ARMA model depends on the results of the V/U/M/N/S classification as shown in ref [4]. This algorithm can be implemented as follows:

(1) Initialize the values of P1, $\hat{\theta}_1$, λ_{min} and E1 and specify the filter order. (Experience shows that the values of P1 and E1 are insensitive to the algorithm provided they are adequately large, (e.g., $P_1 = 10^2$, $E_1 = 10^6$)).

(2) Compute the filter gain K_k, error covariance P_k and prediction error e_k using the WRLS-VFF algorithm as in Table 1.

(3) Compute λ_k . If $\lambda_k < \lambda_{min}$, then $\lambda_k = \lambda_{min}$, where λ_{min} is given by equation (6)

(4) Calculate the new filter coefficient vector $\hat{\theta}_k$ as in Table 1.

(5) Check for glottal closed phase using VFF based in ref [4]. If there is a closed glottal interval, then extract the formants and their bandwidths by solving for the roots of the polynomial obtained from $\hat{\theta}_k$.

(6) Go to (2) until end of data.

III. Performance Evaluation of the Proposed algorithm

The WRLS-VFF-VT algorithm described in section 2 is implemented and compared to block data processing algorithms : two-channel Closed Phase Covariance (CPC)[5].

Synthetic speech is used for our initial performance evaluation because we can specify and control such parameters as the formants, their bandwidths, and the excitation source with a formant synthesizer[6]. The isolated words and sentences spoken by a male subject is used to illustrate the performance of different algorithm such as two-channel CPC. The performance of each algorithm is evaluated according to its formant tracking ability and its formant bandwidth estimation. For formant/ bandwidth extraction, the roots of the numerator and denominator polynomials of the ARMA model are determined for the glottal closed phase interval for each period. All roots with a Q factor (center frequency divided by bandwidth) less than one were eliminated. The remaining roots are retained as potential formant roots. No other form of formant trajectory smoothing or filtering is done.

The analysis conditions for each method are indicated below.

1) For the AR model a 12th order was used.
2) For the ARMA model p=12 and q=6.
3) No preemphasis was used, since the closed phase vocal tract filters derived with and without preemphasis were virtually the same.

3.1. The generation of synthetic speech signal

We illustrate our results here with several synthetic speech signals generated by using a formant synthesizer[6] for the isolated nasal/nonnasal words and the sentence.

For synthesizing sustained phonation, such as sustained vowel /a/and nasal sounds /m/, /n/, the amplitude, fundamental frequencies, and first four formant frequencies and bandwidths were used as constants as shown in Table 2. The frame size was also kept constant by 100 points (10ms interval).

For the synthesizing of the all-voiced multiple sounds utterance, "We were away a year ago", the values of the above parameters are variable. The frame size is equal to the pitch-period. The formant frequencies and bandwidths are estimated from pitch synchronous LPC analysis and were smoothed by hand with WAVES + DSP package.

Table 2. Formant (F)/ antiformant(FZ) and bandwidth (B) used for generating synthetic speech signals (all frequencies are in Hz).

signal	F1	F2	F3	F4	B1	B2	B3	B4	FZ1	BZ1
/m/	390	1250	2150	3150	60	150	200	300	780	80
/n/	390	1250	2650	3950	60	150	200	300	1780	600
/a/	508	1069	2626	3035	102	48	131	213		

3.2. erimental results

The estimated formant frequencies and bandwidths for synthetic speech /a/ using the WRLS-VFF-VT and the two-channel CPC are shown in Figure 2. The analysis results from the two-channel CPC [Figure 2 (a) (c)] show that the bandwidths of formants cannot be accurately estimated due to the influence of input pulse, but the formant frequencies can be estimated accurately. Compared with the pitch-synchronous CPC method, the WRLS-VFF-VT can estimate the formant frequencies and bandwidths of the reference model more accurately [Figure 2(b) (d)]. For synthetic speech of /m/, /n/ also show the similar results.

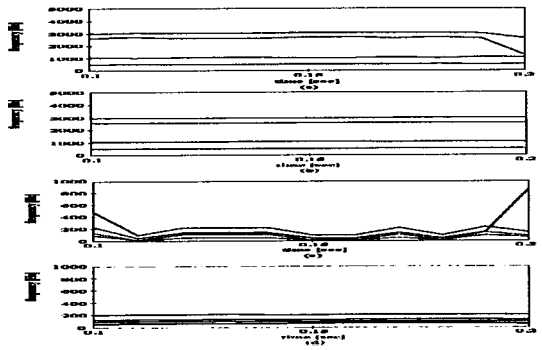

Figure 2. Formant frequency and BW tracks for synthetic speech: (a)(c) 2 channel CPC, (b)(d) proposed method.

The analysis results for synthetic utterance, "We were away a year ago", are illustrated in Figure 3. A time varying reference model which has four formants as shown in Figure 3(a) is used for the formant synthesizer. Figure 3(b) illustrates the analysis result of the synthetic speech signal of this reference model. We see that formants are estimated accurately

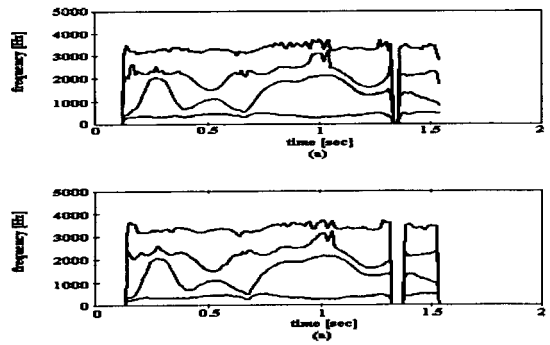

Figure 3. Formant tracking for synthetic speech (a) original formant, (b) estimated formant.

Figure 4 shows the formant tracking contours for the real speech of the utterance "We were away a year ago", estimated by the two-channel CPC and WRLS-VFF-VT methods, respectively. These data are for a male voice. An examination of the formant contours of this figure shows that the WRLS-VFF-VT method is better than the two-channel CPC in terms of tracking ability and smoothness of the tracks of formant estimates. The two-channel CPC methods introduces abrupt discontinuities in the formant frequencies. The results show that the WRLS-VFF-VT method give the better results for the formant trajectories, in the sense that there are very few sudden jumps in the contours.

From the experimental results, we found that the block data processing techniques, such as the two-channel CPC, gave reasonable estimates of the formant/antiformant and their bandwidths. However, the data windows used by these methods included the effects of the periodic excitation pulses, which affected the accuracy of the estimated formants and their bandwidths. On the other hand the WRLS-VFF-VT methods, which eliminated the influence of the pulse excitation by using an input estimation as part of the algorithm, gave very accurate formant/bandwidth estimates and good spectral matching.

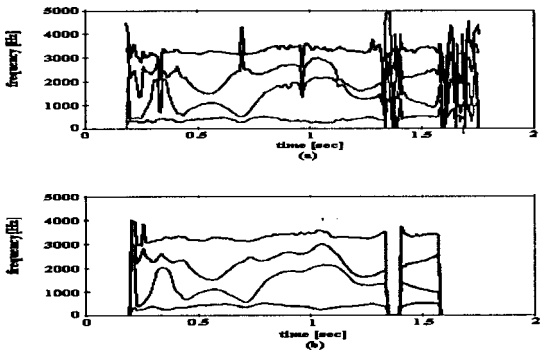

Figure 4. Formant tracking for real speech (a) two-channel CPC, (b) proposed algorithm.

IV. Conclusion

The proposed algorithm has been tested extensively on considerable speech data. Our results indicate that the closed phase WRLS-VFF-VT method of analysis seems superior to other block data processing techniques, such as the two-channel CPC method. The proposed method can be used for all-pole model analysis (e.g., vowels and diphthongs) as well as for pole-zero analysis (e.g., fricatives and nasals). The CPC method, which some

well when the closed phase interval is short, as with female or children's voices, or when the vocal tract characteristics change rapidly, as with the vowel/-consonant transitions and some glide sounds. This is attributed to poorly estimated vocal track filter parameters obtained from a covariance matrix using a short data interval. Furthermore, matrix ill-conditioning problems may occur when solving the least squares equation in the CPC method. The adaptive recursive algorithms derived from a least square cost function are known to converge rapidly(for short data records)[7], and have an excellent capability to "track" an unknown parameter vector.

In the proposed WRLS-VFF-VT algorithm, a variable forgetting factor is used to allow the estimation process to track the time-varying parameters even more quickly. Finally comparing with WRLS-VFF of fixed threshold which was suggested by Ting[8], the proposed variable threshold algorithm shows superiority in both the phoneme boundary and the voiced/unvoiced boundary.

References

1. Liporace, L. A., "Linear Estimation of non- stationary signals," Journal of Acoustical Society of America, vol. 58, pp. 1288-1295, 1975.
2. Casscuberta, F. and Vidal, E. "A nonstationary model for the analysis of transient speech signals," IEEE Trans. on ASSP, vol. 35, pp. 226-228, 1987.
3. Grenier, Y., "Time-dependent ARMA modelling of nonstationary signals," IEEE Trans. on ASSP, vol. 31, pp. 899-911, 1983.
4. Lee, K. S. "Pitch synchronous analysis/synthesis system using WRLS-VFF-VT algorithm," Ph.D Thesis, Univ. of Florida, 1992.
5. Krishnamurthy, A. K. and Childers, D. G., "Two-channel speech analysis," IEEE Trans. on ASSP, vol. 34, no. 4, pp. 730-743, 1986.
6. Klatt D.H., "Software for a cascade/parallel formant synthesizer," Journal of Acoustic Society of America, vol. 67, no. 3, pp. 971-995.
7. Haykin, S., "Adaptive Filter Theory," Prentice-Hall, Inc., Englewood Cliffs, NJ, 1985.
8. Ting, Y. T., "Adaptive estimation of time-varying signal parameters with applications to speech," Ph.D Thesis, Univ. of Florida. 1989.

▲ Kyo-Sik Lee

The Journal of the Acoustical Society of Korea, Vol. 17 No. 3E, 1998.

▲ Kyu-Sik Park

The Journal of the Acoustical Society of Korea, Vol. 18 No. 2E, 1999.