NONLINEAR PREDICTIVE ANALYSIS OF SPEECH BY ITERATIVE APPROACH

Hirobumi Tanaka and Tetsuya Shimamura

Graduate School of Science and Engineering, Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama 338-8570, Japan
phone: +81 48 858 3776, fax: +81 48 858 3776, email: {tanahiro,shima}@sie.ics.saitama-u.ac.jp

ABSTRACT

The filter involving the adaptation scheme of Volterra Series Least Mean Square(VSLMS) algorithm is a representative adaptive nonlinear filter, which has been applied to a lot of engineering applications. However, when the VSLMS filter is used as an adaptive predictor of speech, a large number of speech data samples are required to minimize the predictive error. And if the VSLMS predictor is used for short-term prediction with a high order of the quadratic kernel to increase the predictive gain, it is suffered from its numerical unstability. To conquer such problems, an iterative approach is proposed in this paper. The iterative approach gives an effect to utilize a large number of speech data samples by using a segmented speech signal repeatedly. Experiments are conducted on continuous speech and it is shown that the predictive accuracy of the VSLMS predictor is improved by relying on the iterative approach.

1. INTRODUCTION

Speech production is extensively assumed to be modeled by the use of a linear filter. In fact, the technique of linear prediction (LP)[1] has been used in many speech processing systems, in which the speech signal is modeled as the output of a linear all-pole filter whose input is white noise for unvoiced speech or a chain of impulses for voiced speech.

When the LP analysis of a speech signal is made, basically the coefficient vector results in an accurate representation of the speech signal if the predictive order is determined adequately. However, the LP analysis on voiced speech sometimes may lead to an inaccurate result. This is because the excitation sequence of voiced speech has impulsive characteristics and affects adversely the performance of the LP analysis[2]. In particular, this is visualized in the residual sequence produced by the LP analysis[3].

To overcome such a problem, Thyssen et al[6] addressed the use of nonlinear prediction based on multilayer perceptron. On the other hand, recently Vorogle et al[7] reported a new configration of predictive analysis relying on recurent neural networks. However, in neural networks, we generally need experimental references(knowledge) in order to determine their layer structures. Furthermore, a large number of data samples are needed to achieve convergence and the corresponding desired solution cannot be always obtained.

The above-mentioned methods are implemented in a form of batch processing, but a sequential form, adaptive nonlinear prediction, is also known[4]. Mumolo et al[5] proposed an adaptive nonlinear predictor based on a configration of Volterra Series(VS) filter for the purpose of ADPCM and achieved improvement of a performance relative to the counterpart of its linear predictor. If the LMS algorithm[8] is deployed for the adaptation procedure of the VS filter, the computation of the resulting VSLMS predictor is very simple. However, the convergence speed of the VSLMS predictor is not sufficiently fast. This results in an inaccurate representation of speech, increasing the predictive error. Although Carlos et al[9] applied the Kalman filter theory to the VS fiter, the resulting adaptive filter is too complicated for implementation, which seems to be inadequate for speech applications.

In this paper, we address a technique that keeps the computational simplicity of the VSLMS predictor and provides an accurate representation of speech. An iterative operation is used, with which the VSLMS predictor is implemented.

2. VSLMS PREDICTOR

In this section, we consider the quadratic VSLMS predictor as depicted in Figure 1, in which the speech signal $s(n)$ is assumed to be predicted from its previous values such as

$$\hat{s}(n) = \phi(s(n-1), s(n-2), ..., s(n-M)) \qquad (1)$$

where the hat denotes an estimate and $\phi(\cdot)$ means a mapping function including adjustable parameters.

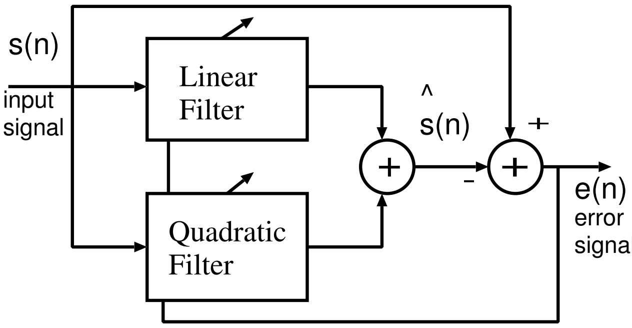

Figure 1: Quadratic VSLMS predictor

When the function $\phi(\cdot)$ in (1) behaves nonlinearly, an adaptive nonlinear predictor can be realized. Mumolo et al.[5] deployed the quadratic Volterra filter for the function $\phi(\cdot)$. In this case, the estimate of the speech signal is given by

$$\hat{s}(n) = \sum_{i=1}^{M_1} a_i s(n-i) + \sum_{i=1}^{M_2}\sum_{j=i}^{M_2} b_{ij} s(n-i) s(n-j) \qquad (2)$$

where $a_i, i = 1,2,...,M_1$ and $b_{ij}, i, j = 1,2,...,M_2$ correspond to the linear and quadratic predictive coefficients, respectively. In (2), the symmetrical charactaristic of quadratic predictive coefficients, $b_{ij}(n) = b_{ji}(n)$, is considered and the redundancy in quadratic configration is omitted.

The adaptation procedure for the VSLMS predictor is given by

$$e(n) = s(n) - \mathbf{q}(n)^T \mathbf{h}(n) \tag{3}$$

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \frac{\mu}{\mathbf{q}(n)^T \mathbf{q}(n) + \beta} \mathbf{q}(n) e(n) \tag{4}$$

where $\mathbf{q}(n)$ is the input vector at time n

$$\begin{aligned} \mathbf{q}(n) = [&s(n-1), s(n-2), ..., s(n-M_1), \\ &s(n-1)^2, s(n-1)s(n-2), ..., s(n-M_2)^2]^T \end{aligned} \tag{5}$$

and $\mathbf{h}(n)$ is the coefficient vector at time n

$$\mathbf{h}(n) = [a_1(n), a_2(n), ..., a_{M_1}(n), b_{11}, ..., b_{M_2 M_2}]^T. \tag{6}$$

The μ and β mean the step size and stabilized parameters for the normalized LMS algorithm, respectively.

The VSLMS predictor can minimize the mean square predictive error. Hence the least square solution for the predictive coefficients can be expressed analytically. Since the nonlinearlity of speech is deeply related with the quadratic kernel for a VS filter[6], the VSLMS predictor can result in an excellent nonlinear predictor reducing more the predictive error than the linear predictor. However, the VSLMS predictor requires a large number of data samples to minimize the predictive error. Additionally, a large number of coefficients are required to be set for the quadratic kernel. This is realized by increasing the order of the quadratic kernel. However, the VSLMS predictor with a high order for the quadratic kernel very often become numerically unstable on real speech. Thus, in ths paper, we set out to suppress the appearance of the unstable phenomenon of the VSLMS predictor by keeping the order of the quadratic kernel comparatively lower.

3. ITERATIVE APPROACH

For the purpose of speech analysis, the convergence speed of the adaptive predictor to be deployed should be carefully considered. It is said that on real countinuous speech, the stationary property is kept during 20-30 *ms*. For such a short length of speech, the convergence of any adaptive predictors may not be guaranteed. Reducing the number of data samples for the purpose of predictive analysis means increasing the predictive error. The minimum predictive error is obtained when the convergence of the deployed adaptive predictor is achieved. Therefore, there exits a trade-off between the number of data samples and the preditive accuracy.

To solve this problem, we propose an iterative approach in which the speech data in the analysis frame are used repeatedly. The method is implemented as follows.

1. The initial values of the predictive coefficient vector are set to zeros (for the VSLMS predictor, $\mathbf{h}(1) = \mathbf{0}$).
2. The adaptation is carried out in the analysis frame from 1 to N.
3. The final coefficient vector and input vector (for the VSLMS predictor $\mathbf{h}(N)$ and $\mathbf{q}(N)$) are stored.
4. As the initial setting of the coefficient vector and input vector, those obtained in Step 3 are used.
5. The value of step size is decreased by a fixed value of p, and go to Step 2.

With a small number of data samples, the predictive analysis of speech may not be conducted adequately. However, if the predictive coefficient vector and input vector are stored as the initial setting for the adaptation, then those are used in the same analysis frame and we can make the predictive analysis again, which corresponds to the case where the analysis frame length is extended by a factor of two. This operation is easily extended. If the framed data samples are repeatedly used, then a large number of data samples could be utilized for the predicitve analysis. In the next section, experiments are conducted on real speech to confirm the validity of the VSLMS predictor with the iterative technique.

4. EXPERIMENTS

To investigate the performances of the VSLMS predictor with the iterative technique, we conducted experiments on real speech signals sampled with 10 kHz.

4.1 Relation between the number of iterations and predictive gain

At first, we investigated the relation between the number of iterations and predictive gain(SNR). In the experiments, speech data used are 2 male and 2 female speakers, each of which consists of 5 vowels. We used only 300 data samples for each vowel here by considering the stationary property of continuous speech. The predictive order and stabilized parameter were commonly set to $M_1 = 10$, $M_2 = 3$ and $\beta = 0.05$, these gave the stability of the VSLMS predictor. At the first iteration, the step size was set to the optimum value the VSLMS predictor provides on each vowel (some preliminary experiments were conducted to find the optimum value). And then, as the number of iterations was increased, the value of step size was decreased by 0.05 at each iteration. The predictive gain was evaluated as

$$SNR(dB) = 10\log_{10} \frac{\sum_{n=n_1}^{n_2} s_w(n)^2}{\sum_{n=n_1}^{n_2} e_w(n)^2} \tag{7}$$

with the setting of $n_1 = 1$ and $n_2 = 300$, where $s_w(n)$ is the framed speech signal and $e_w(n)$ is the corresponding predictive error.

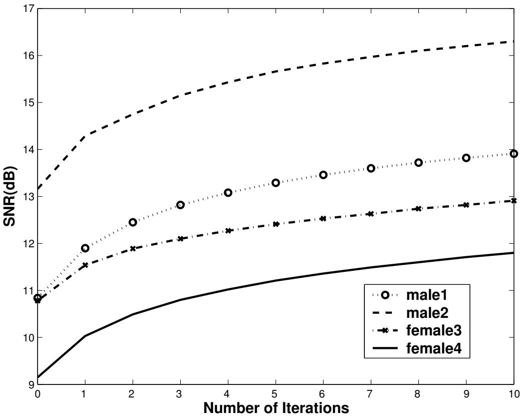

Figure 2: Relation between the number of iterations and the average of predictive gains on 5 vowels for the VSLMS predictor with iterative approach

Figure 2 shows the relation between the number of iterations and the average of predictive gains on 5 vowels. From this figure, we notice that the predictive gain is drastically improved only by 1 iteration, and after about 10 iterations, one reaches to the convergence providing an improvement of 2-3 dB. Figure 2 obviously validates the effectiveness of the iterative approach.

4.2 Experiments on continuous speech

Next, we conducted experiments on continuous speech. Speakers are 2 male and 2 female Japanese. Each sentence is about 5 s. These speech data were adopted from "Multi-Lingual Speech Database for Telephonometry 1994(NTT Advanced Technology Corp.)."

Both of the LMS predictor and VSLMS predictor are investigated. The predictive order and stabilized parameter for both are commonly set to $M = 10$ and $\beta = 0.05$. The quadratic predictive order for the VSLMS predictor is set to $M_2 = 3$ to keep the VSLMS predictor stable. The step size for both predictors is optimized to achieve the best performance. The frame length is 30 ms. As a criterion of the performance, we used the predictive gain in (7) with the setting of $n_1 = 1$ and $n_2 = 300$.

For continuous speech, the relation between the analysis frame length and the speech period should be considered more carefully. Thus, we investigated additionally two different frame lengths ; 24 ms and 36 ms. (In these cases, we evaluated the predicitive gains with the setting of $n_1 = 1$ and $n_2 = 240$, and with the setting of $n_1 = 1$ and $n_2 = 360$, respectively).

Tables 1 and 2 show the resulting predictive gains on continuous speech uttered by male and female speakers, respectively. In these Tables, "adaptive" means that the adaptive predictor is adaptively implemented on all the data samples of 5 s. Tables 1 and 2 show that for the VSLMS predictor, the frame based processing (the non-iterative processing) provides better performance than the adaptive processing regardless if the iterative operation is used or not. Futhermore, the iterative processing provides better performance than the non-iterative processing. From these results, we deduce that the iterative approach could improve the accuracy of adaptive prediction on continuous speech.

4.3 Application of iterative approach

In this subsection, a proposal system is introduced for effectively making use of the iterative approach. Figure 3 shows a configuration of the predictive analysis scheme to improve the performance further. This scheme is derived by considering that an improvement in predictive gain is obtained by utilizing the speech samples predicted in the regions where the predictive coefficients are converged. At first, plural predictors are prepared in parallel (these make a different arrangement for calculating the predictive gains. In Figure 3, the case where 3 predictors are prepared is shown). On continuous speech, one fame length for the iterative processing is selected so that the stationary property of speech is maintained (in the experiments, it is set to 300 samples(30 ms)). Speech data samples are inputted to the first predictor, and the iterative processing is performed. And the speech data predicted in the region where the predictive coefficients are converged are outputted (In Figure 3, we determined the region where the predictive coefficients are converged is from 200 to 300 samples(from 20 to 30 ms). Dotted circles in Figure 3 mean the regions).

The second predictor is set up so that the speech data samples delayed by 100 samples(10 ms) are inputted. After that, the iterative processing is performed and the speech data predicted in the region where the predictive coefficients are converged are outputted.

The third predictor is set up so that the speech data samples delayed by 200 samples(20 ms) are inputted. After that, the processing of the third predictor is similar to those of the first and the second predictors.

For the first predictor, the next speech data samples, which are delayed by 300 samples(30 ms) from the starting point, are inputted. Serial processing like this is conducted. In such a way, we could utilize the iterative approach more effectively.

We conducted experiments based on the configuration in Figure 3. The parameters used in the experiments are similar to those in Subsection 4.2. Table 3 shows the average of the predictive gains in (7) with the setting of $n_1 = 201$ and $n_2 = 300$ where all the frames of each sentence have been evaluated for the average. Comparing the predictive gains obtained by the conventional methods(adaptive LMS and adaptive VSLMS) in Tables 1 and 2 with that in Table 3, we see that an improvement of 4-5 dB is obtained. This means that we can improve the predictive accuracy by using the speech data samples predicted in the region where the predictive error is converged.

5. CONCLUSION

In this paper, nonlinear predictive analysis for speech by iterative approach was proposed. As the predictor, the VSLMS adaptive filter was selected. For the purpose of improving the predictive accuracy the VSLMS predictor provides, we addressed an iterative method. The validity of the iterative method was confirmed by experiments on real speech.

Furthermore, to make use of the iterative method more effectively, we considered the region where the predictive coefficients are converged, and a configuration of parallel structures of the VSLMS predictors was derived. As a result, it was confirmed that the predictive accuracy was further improved.

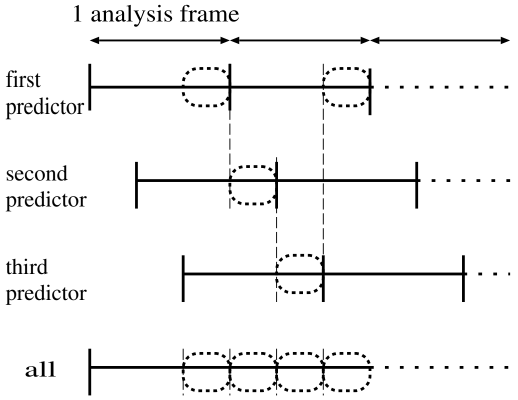

Figure 3: Configuration of parallel VSLMS predictors

REFERENCES

[1] B.S.Atal and S.Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acous-

Table 1: SNR on continuous speech uttered by male speakers

		frame length(*ms*)	24	30	36	average
sentence1	LMS	adaptive(conventional)	–	–	–	11.46
	VSLMS	adaptive(conventional)	–	–	–	11.72
		no iterations	12.98	13.46	13.36	13.27
		5 iterations	13.93	14.37	14.50	14.27
		10 iterations	14.80	14.72	15.10	14.87
sentence2	LMS	adaptive(conventional)	–	–	–	11.31
	VSLMS	adaptive(conventional)	–	–	–	11.69
		no iterations	12.72	13.01	12.83	12.85
		5 iterations	13.46	14.26	14.17	13.96
		10 iterations	14.52	15.35	15.08	14.98

Table 2: SNR on continuous speech uttered by female speakers

		frame length(*ms*)	24	30	36	average
sentence1	LMS	adaptive(conventional)	–	–	–	12.54
	VSLMS	adaptive(conventional)	–	–	–	12.65
		no iterations	14.99	15.45	15.90	15.45
		5 iterations	16.18	16.01	16.28	16.16
		10 iterations	16.38	16.29	16.68	16.45
sentence2	LMS	adaptive(conventional)	–	–	–	12.38
	VSLMS	adaptive(conventional)	–	–	–	12.58
		no iterations	14.38	14.53	14.90	14.60
		5 iterations	15.29	15.39	15.30	15.33
		10 iterations	14.88	16.43	15.38	15.56

Table 3: SNR evaluated by the configuration in Figure 3

		speaker	male	female
VSLMS	sentence1	5 iterations	16.67	16.79
		10 iterations	16.98	19.24
	sentence2	5 iterations	17.29	17.23
		10 iterations	19.35	17.56

tics Society of America, Vol.50, No.2, pp.637-655, 1971.

[2] S.Chandra and W.C.Lin, "Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis", IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-22, No.6, pp.403-415, 1974.

[3] L.R.Rabiner, B.S.Atal and M.R.Sambur, "LPC prediction error - analysis of its variation with the position of the analysis frame", IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-25, No.5, pp.434-442, 1977.

[4] J.D.Gibson, J.L.Melsa and S.K.Jones, "Digital speech analysis using sequential estimation techniques", IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-23, No.4, pp.362-369, 1975.

[5] E.Mumolo, A.Carini and D.Francescato, "ADPCM with non linear predictors", Proc. EUSIPCO94, pp.387-390, 1994.

[6] J.Thyssen, H.Nielsen and S.D.Hansen, "Non-linear short-term prediction in speech coding", Proc. ICASSP94, pp.I-185-188, 1994.

[7] E.Varoglu and K.Hacioglu, "Recurrent neural network speech predictor based on dynamical systems approach", IEE Proc. -Vis. Image Signal Process, Vol. 147, No2, 2000.

[8] B.Widrow et al., "Stationary and nonstationay learning characteristics of the LMS adaptive filter", Proc.IEEE, vol.64, pp.1151-1162, Aug. 1976.

[9] Carlos E. Davila, Ashley J. Welch and H. Grady Rylander, "A second-order adaptive Volterra filter with rapid convergence", IEEE Trans, Acoustics, Speech and Signal Processing, Vol.ASSP-35, NO.9, pp.1259-1263, 1987.