SPEECH MODELING BY CORRELATION FITTING

Jianguo Huang

Northwestern Polytechnical University

Xian, P.R. China

Leland B. Jackson

University of Rhode Island

Kingston, RI 02881, U.S.A.

Abstract--We apply a previously proposed technique for AR modeling by fitting windowed correlation data to speech modeling in additive white noise. Using the correlation fitting (CF) method, the estimated speech autocorrelation function can be approximated over many more than the minimum number of correlation lags to produce a much more accurate fit to the corresponding power density spectrum, reducing the effect of the noise. Simulation shows the improved results.

Keywords: AR modeling of speech, Correlation fitting mehtod, Linear prediction coding, Yule-Walker equations.

1. Introduction

The method for autoregressive(AR) and autoregressive-moving-average(ARMA) modeling of stationary stochastic signal have previously been proposed based upon fitting the model autocorrelation function to the estimated (and biased) autocorrelation in the least-square sense over many more than the minimum number of autocorrelation value (1,2). Here we apply the correlation fitting(CF) method to the case of speech modeling. For speech analysis the method of linear prediction coding(LPC) is one of the most powerful technique (3). The CF method has the same results as LPC when only the minimum number of autocorrelation lags are used.

But in the case of speech contaminated by white noise the LPC analysis technique degrades rapidly. For high signal-noise-ratio(SNR), greater than 30db, the additive noise only covers very small rigions of the spectrum as well as autocorrelation. Speech signal still is approximately a pth-order all-pole signal. The fitting of the first p+1 autocorrelation ensures fitting the correspoding power density spectrum. When the sources of noise become significant (SNR<30db) the corrupted speech signal is no longer autoregressive. It is autoregressive-moving-average. The spectral noise zeros affect the true AR poles to move towards origin, thus producing a flat spectrum (5). So LPC analysis does not preserve the formant structure of the speech spectrum.

This point is shown in Fig.1 where the autocorrelation and spectrum of speech-like signal segmant corrupted by additive white noise is shown by a solid curve. The order of AR model is p=2. The autocorrelation and spectrum of all-pole model is plotted with the dotted curve. LPC matches the first two lags exactly, but mismatches higher lags and degrades quickly. The corresponding LPC spectrum becomes flat as shown in Fig.1(b).

AR modeling in the presence of noise has been a subject of much research (6-10) in recent years, and various modification to LPC have been proposed to retain high performance. Some of them need more computation to estimate the variance of noise, or use the high-order Yule-Walker equations(HOYWE) which delete the first p+1 equations, i.e. all of the original Yule-Walker equations.

By use of the correlation fitting method for AR modeling of speech we get improved results, especially in the case of speech contaminated by noise. The correlation fitting algorithm is derived in section 2. The robustness of CF is described in section 3. Finally, section 4 shows the simulation results and section 5 gives conclusions.

2. AR Modeling by the Correlation Fitting

The AR correlation fitting algorithm (1,2) is derived in this section.

Assume x(n) is a stationary autoregressive stochastic signal. It can be well represented by an AR model after solving the famous Yule-Walker equations

$$\hat{R}_p \underline{a}_p = \sigma^2 \underline{\delta}_p \qquad (1)$$

where $\underline{a}_p$ is the coefficient vector of a pth-order AR model, and $\hat{R}_p$ is a Hermitian Toeplitz matrix comprised by the p+1 estimated correlation values of x(n), i.e. $\hat{r}(0), \hat{r}(1), \ldots, \hat{r}(p)$ which are matched exactly by the resulting AR model (4).

In order to improve the performance we use the fitting error criterion

$$e(n) = \hat{r}(n)u(n) - w_T(n)r(n)u(n) \qquad (2)$$

with

$$min \sum_{n=0}^{L} e^2(n) \qquad (3)$$

where $\hat{r}(n)$ is the estimated (and biased) autocorrelation of x(n), r(n) is the model autocorrelation, u(n) is the unit step function, and $w_T(n)$ is the triangular window corresponding to the bias in $\hat{r}(n)$.

Utilizing the fact that $h(n) * a(n) = \delta(n)$, we can write (2) as

$$h(n) * [\hat{r}(n)u(n)] * a(n) = w_T(n)r(n)u(n) + e(n) \qquad (4)$$

or in matrix form

$$H\hat{R}'_L \underline{a}_p = W_T \underline{r} + \underline{e}_L \qquad (5)$$

The solution vector is given by

$$\underline{a}_p = (H\hat{R}'_L)^{\sharp} W_T \underline{r} \qquad (6)$$

where $\sharp$ denotes the pseudo-inverse matrix. A close approximate solution can be obtained iteratively via

$$\underline{a}_{p,i+1} = (H_i \hat{R}'_L)^{\sharp} W_T \underline{r}_i \qquad (7)$$

where H_i and $\underline{r}_i$ correspond to the ith estimate $\underline{a}_{p,i}$. The initial estimate $\underline{a}_{p,0}$ is obtained from either the Yule-walker equations or the overdetermined normal equations.

If x(n) is a speech segment, the parameters of AR model of the speech can be well obtained by using CF algorithm.

3. Robustness of CF AR Modeling of Speech

In the case of a speech signal corrupted by additive white noise the estimated autocorrelation function is $\hat{r}(n) + \sigma_n^2\delta(n)$, where σ_n^2 is the noise variance. The zero-lag correlation value is theoretically altered by white noise. From the Yule-Walker equations

$$\begin{bmatrix} \hat{r}(0)+\sigma_n^2 & \hat{r}(-1) & \cdots & \hat{r}(-p) \\ \hat{r}(1) & \hat{r}(0)+\sigma_n^2 & \cdots & \hat{r}(-p+1) \\ \vdots & \vdots & & \vdots \\ \hat{r}(p) & \hat{r}(p-1) & \cdots & \hat{r}(0)+\sigma_n^2 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (8)$$

we can see that the noise term affects every equation. Using these equations which match exactly noisy autocorrelation it is impossible to get correct parameters of AR model. The CF method uses more correlation lags which are not altered by noise, i.e. it adds more equations in solving (5) (L>p) to obtain the parameters of AR model. It equally gives a small weight on the noisy correlation $\hat{r}(0)+\sigma_n^2$ to reduce the effect of the white noise.

Fig. 2 shows the position of speech spectral poles on the Z-plane and illustrates the effect of noise as well as robustness of CF AR modeling. A 12th-order AR model is generated to model a speech segment of 20 msec. The 12 roots from Z polynominal of AR model correspond 12 poles which include 5 pair of complex and two real. The ordinal number of the poles is shown in Fig. 2 by number 1 to 7 for convenient analysis. The dots are the correct position of the poles by the Yule-Walker equations without noise. When the white noise is added (SNR=5db), the power spectral density (PSD) of the noisy speech is

$$P_Y(Z) = \frac{\sigma^2}{A(Z)A(Z^{-1})} + \sigma_n^2 = \frac{\sigma^2 + \sigma_n^2 A(Z)A(Z^{-1})}{A(Z)A(Z^{-1})} = \frac{\sigma_b^2 B(Z)B(Z^{-1})}{A(Z)A(Z^{-1})} \qquad (9)$$

if we let $\sigma_b^2 B(Z)B(Z^{-1}) = \sigma^2 + \sigma_n^2 A(Z)A(Z^{-1})$.

Thus, the noise has introduced zeros into the PSD of noisy speech which are located in between the true AR poles and the origin. It causes the AR poles to move towards the origin, thus generating a smoothing PSD shown in Fig. 2 for the poles 1, 2, 4, 5 which are near the unit circle corresponding to the four formant peaks. The symbol "x" denotes the pole of the noisy speech using the YW equations.

In solving the Eq. (5) the CF method uses more correlation lags (L=25) to get more accurate AR model, the poles of which (denoted by "△") are close to the correct position of poles. This improvement is explained as follows. More correlation lags to be used correspondingly add a small weight on noise variance σ_n^2 in (7). It makes the zeros introduced by the noise move towards the origin and far from the true poles to reduce the effect of the noise on the true poles. Hence, the poles of the new AR model estimated by the CF go to the position of the true poles. But the high frequency pole 5 is far from the true pole because the noise is much stronger than the speech.

4. Simulation Results

Simulation experiments are presented in this section. The experiments use the true speech segment of 20 msec sampled at 8 kHz. The order of AR model is p=12 either for YW using the minimum lags, or for CF method with L=40 lags. The autocorrelation of YW and CF are shown in Fig. 3(a) compared with the autocorrelation of the speech segment denoted by the solid line using the same speech segment as in Fig. 2. As expected, the first 12 lags are exactly matched by YW, but the agreement for the high lags is very poor due to the additive white noise (SNR=5db). Although the CF method only matches approximately the first 12 lags, but the agreement for the high autocorrelation lags which are not altered by the white noise is much better than the YW. As compared with the true autocorrelation of the speech without the noise the CF is much closer than YW shown in Fig. 4(a) and the high lags become smaller due to the noise. The comparison of the power spectral envelope of the speech with noise by using YW and CF is shown in Fig. 3(b) and that without noise is shown in Fig. 4(b). The spectrum of YW is smoothed by the noise and larger errors occur in the frequency estimation of the formants, in which less error is in the first two formants, and more in another two formants. The spectrum of CF is better

than that of YW to retain more accurate frequency estimation of the first three formants. This is important for speech modeling. For the spectral peak of the fourth forment, both of them are obviously affected by noise, because noise has bigger energe than speech in the high frequency rigions. These are consistent with the analysis in section 3 and also shown in Fig. 2. Table 1 lists the number of iterations versus the lags in which the same speech segment is used as in the first experiment. It shows that the more the lags are used, the less the number of iterations are needed.

5. Conclusions

We have developed an extension of AR modeling by fitting windowed correlation data to speech. Unlike either the YW equations we use many more than the minimum lags, or the overdetermined method we minimize the fitting error in the least-square sense to produce a much more accurate fit to the correct power density spectrum. In the present of noise the CF method shows the robust characteristics in reducing the effect of the noise and obtains more accurate spectral estimation for the formants of the speech. The simulation illustrates the improved performance by the CF method.

References

(1) L. B. Jackson, "AR modeling by least squares fitting of windowed correlation data," Proc. of ASSP Workshop on Spectral Estimation, Boston, Nov. 1986, pp. 67-69.

(2) L. B. Jackson, Jianguo Huang and K. P. Richards, "AR, ARMA, and AR-in-noise modeling by fitting windowed correlation data," Proc. ICASSP, Apr. 1987, pp. 2039-2042.

(3) L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals (Chapter 8)," Prentice-Hall, 1978.

(4) L. B. Jackson, "Digital Filters and Signal Processing (Chapter 10)," Boston: Kluwer Academic Publishers, 1986.

(5) S. M. Kay, "The effects of noise on the autoregressive spectral estimator, IEEE Trans. on ASSP, Vol. ASSP-27, No. 5, Oct. 1979.

(6) V. K. Jain and B. S. Atal, "Robust LPC analysis of speech by extended correlation matching," Proc. ICASSP, March, 1985, pp. 473-476.

(7) V. K. Jain and B. L. Xu, "Autocorrelation distortion function for improved AR modeling," Proc. ICASSP, Apr. 1987, pp. 356-359.

(8) D. W. Tufts and R. Kumaresan, "Singular value decomposition and improved frequency estimation using linear prediction, IEEE Trans. on ASSP, Vol. ASSP-30, No. 4, 1982, pp. 671-675.

(9) K. K. Paliwal, "Estimation of noise variance from the noise signal and its application in speech enhancement," Proc. ICASSP, Apr. 1987, pp. 297-300.

(10) Y. T. Chan and R. P. Langford, "Spectral estimation via the high order Yule_Walker equetions," IEEE Trans. on ASSP, Vol. ASSP-30, Oct. 1982, pp. 833-840.

Lags	12	13	20	30	40
Num. of Iterations	1	11	10	5	5

Table 1: The number of iterations vs. correlation lags using CF.

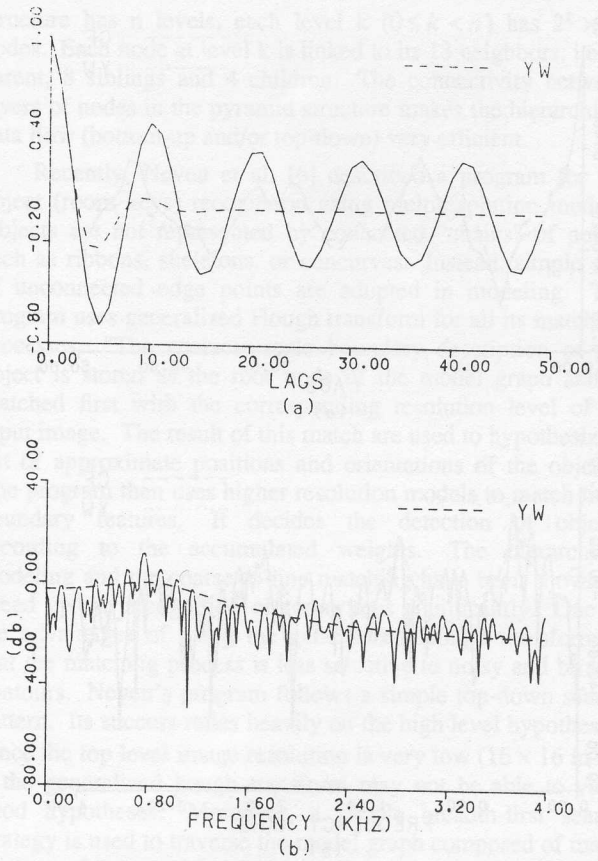

Fig. 1 (a) Autocorrelation function and (b) spectral envelope of the speech-like signal. SNR=10db.

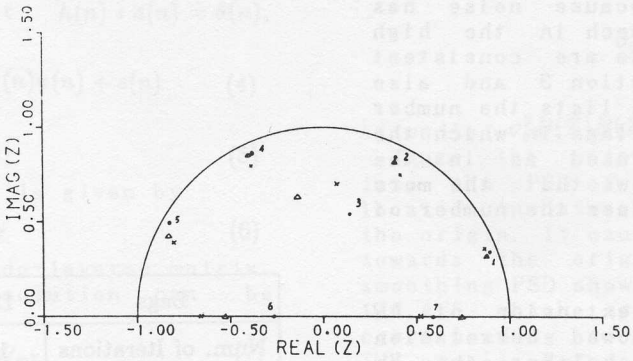

Fig. 2 Pole plot (P=12).
"•" - True pole (by YW).
"×" - Pole by YW with noise, SNR=10db.
"△" - Pole by CF with noise, SNR=10db and L=25.

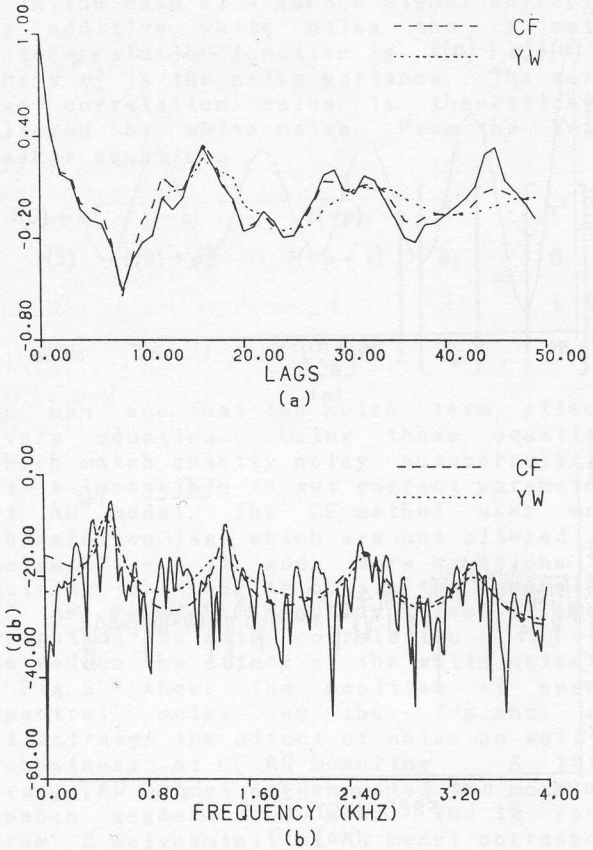

Fig. 3 (a) Autocorrelation function and (b) spectral envelope of the speech segment contaminented by white noise, SNR=5db.

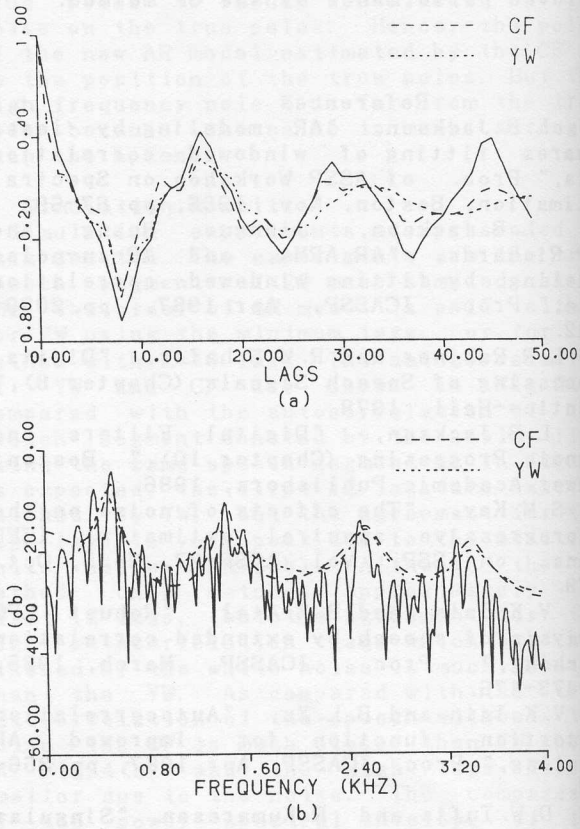

Fig. 4 (a) Autocorrelation function and (b) spectral envelope used in Fig.3 compared with those of speech segment without noise.