On Robust Linear Prediction of Speech

CHIN-HUI LEE, MEMBER, IEEE

Abstract—**In this paper, a robust linear prediction algorithm is proposed. Rather than minimizing the sum of squared residuals as in the conventional linear prediction precedures, the robust LP procedure minimizes the sum of appropriately weighted residuals. The weight is a function of the prediction residual, and the cost function is selected to give more weight to the bulk of smaller residuals while down-weighting the small portion of large residuals. In contrast, the conventional LP procedure weights all prediction residuals equally. The proposed algorithm takes into account the non-Gaussian nature of the excitations for voiced speech. Compared to the conventional LP algorithms, this robust formulation will give a more efficient (less variance) and less biased estimate for the prediction coefficients. When applied to the problem of estimating the center frequency and bandwidth of speech formats, based on finding the roots of the prediction polynomial, the robust LP algorithm therefore achieves more accurate results than the conventional LP procedures.**

The robust LP algorithm can be used in the front-end feature extractor for a speech recognition system and as an analyzer for a speech coding system. Testing on synthetic vowel data demonstrates that the robust LP procedure is able to reduce the formant and bandwidth error rate by more than an order of magnitude compared to the conventional LP procedures. Preliminary experiments on natural speech data indicate that the robust LP procedure is relatively insensitive to the placement of the LPC analysis window and to the value of the pitch period, for a given section of speech signal.

I. Introduction

The linear prediction (LP) model for speech analysis and synthesis was first introduced by Saito and Itakura [1] and Atal and Schroeder [2]. Since that time, the linear prediction formulation has been adopted successfully in various speech applications. The formulation is based on the linear model of speech production [3], [4] defined as

$$S(z) = E(z)\, G(z)\, V(z)\, L(z) \tag{1.1}$$

where $E(z)$ is the driving function to the glottal shaping model $G(z)$, $L(z)$ is the lip radiation model, and $V(z)$ is an all-pole vocal tract model. The combination of $G(z)$, $L(z)$, and $V(z)$ can be approximated by an all-zero filter $A(z)$, called an *inverse filter*, of the form

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i} \approx \frac{1}{G(z)\, V(z)\, L(z)}. \tag{1.2}$$

Equations (1.1) and (1.2) together give an analysis model $E(z) = S(z)\, A(z)$, which can be equivalently expressed in the sampled data domain as

$$s_n + \sum_{i=1}^{p} a_i s_{n-i} = e_n. \tag{1.3}$$

The above formulation indicates that $\{s_n\}$ is an autoregressive model of order p, AR(p). The driving source $\{e_n\}$ is also called the *innovations* process for the AR(p) model. In the context of speech analysis, the innovation is often approximated as either an impulse train with period P for voiced sounds, or as random noise having a flat spectrum for unvoiced sounds.

In the conventional LP (CLP) speech analysis, the linear prediction coefficients a_i's are determined by either the autocorrelation method or the covariance method [5], [6]. The basic formulation of the CLP analysis seeks to find an optimal *fit* to the envelope of the speech spectrum, given a sequence of speech observations $\{s_1, \cdots, s_N\}$. The *fit* is usually obtained by solving for the prediction coefficients a_i's that minimize the sum of squares of the prediction residuals, and therefore, the result is a *least squares* (LS) *fit*. However, the structure of the source excitation is often not taken into account in the LS fitting procedure. It is well known that, for voiced speech, the source is of a quasi-periodic nature with spiky excitations. Those impulsive periodic innovations will often couple with the filter $A(z)$ to produce spurious spectral peaks in the signal spectrum and result in difficulties for LP analysis of high-pitched voices. If this source characteristic can be taken into account for the estimation of the prediction coefficients, a better source/filter separation may be achieved.

In this paper, a robust linear prediction (RBLP) algorithm is proposed. The RBLP algorithm can be used as an analyzer for a speech coding system and as a front-end feature extractor for a speech recognition system. The procedure takes into account the non-Gaussian nature of the source excitation for voiced speech by assuming that the innovation is from a mixture distribution, such that a large portion of the excitations are from a normal distribution with a very small variance while a small portion of the glottal excitations are from an unknown distribution with a much bigger variance. A distribution, with such a property that the tails of the distribution are heavier than those of a nominal Gaussian distribution, is often called *heavy-tailed non-Gaussian*. Mixture (or contaminated) distributions often arise when modeling real-life data [7]. A special type of mixture Gaussian distribution has been used successfully in speech recognition to model the state observation densities of a hidden Markov model (HMM), where the multimode densities cannot be easily characterized by a single multivariate Gaussian distribution [8], [9].

Based on the above mixture source assumption, the

Manuscript received May 7, 1987; revised November 10, 1987.
The author is with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 8719427.

RBLP procedure minimizes the sum of weighted residuals, rather than minimizing the sum of squared residuals. The assigned weight is a function of the prediction residual and the cost function can be selected to assign more weight to the bulk of small residuals while down-weighting the small portion of large residuals. In contrast, the CLP procedure weights all the prediction residuals equally. Based on Statistical Robustness Theory, the proposed algorithm will result in a more efficient (lower variance) estimate for the LP coefficients if the source excitation is sufficiently heavy-tailed non-Gaussian. Therefore, the formant and bandwidth estimates based on solving the prediction polynomial will have smaller variance compared to the results obtained from the CLP analysis. Furthermore, since the RBLP procedure attempts to decouple the source excitation from the vocal tract filter by fitting an AR model to the vocal tract filter, rather than approximating the spectral envelope of the speech signal, a better source and filter separation is expected if the speech signal is well approximated by an AR process with heavy-tailed non-Gaussian excitations.

This paper is organized as follows. In Section II, some background on Robustness Theory is presented. The robust LP algorithm is then derived in Section III. Algorithmic issues related to speech analysis will be discussed in Section IV. Performance of the algorithm on synthetic and natural speech data is evaluated in Section V, and finally, a discussion and concluding remarks are given in Section VI.

II. Review of Robustness Theory

The method of least squares (LS) and its derivatives have been used for many years. However, it is well known that *outliers* have unusually large influence on the resulting LS estimators. The outliers could come from bad data points due to errors in measurement, or simply arise from the nature of the physical data, that is the underlying distribution is not represented by a simple Gaussian distribution. Since outliers can distort the LS *fit* a great deal, the resulting residuals are often misleading because they look more like the normal ones. Accordingly, robust procedures have been created to modify the least squares schemes in order to down-weight the influence of outliers on the final estimators. The best known of the robust procedures is the class of maximum-likelihood-type estimates, or simply *M-estimates*. Other robust estimators, like *L-estimates* and *R-estimates*, also arise in various applications. In this paper, we will only focus our discussion on the *M*-estimates. For more details about Robustness Theory, the reader is referred to Huber [10].

The theory of *M*-estimation was first developed by Huber [11] for the estimation of the location and scale parameters of a sequence of independent and identically distributed (i.i.d.) observations. It has then been successfully generalized for robust hypothesis testing [12], for robust regressions [13], and for robust estimation of the parameters of autoregressive (AR) models [14], and of autoregressive and moving-average (ARMA) models [15]. The essence of the *M*-estimation procedure will be illustrated in the following discussion on location parameter estimation. Given a sequence of observations $\{x_1 \cdots x_N\}$, such that $x_i = \mu + e_i$, the problem is to estimate the location parameter μ. The distribution of the errors e_i's is not assumed to be known exactly, the only assumptions are that $\{e_1 \cdots e_N\}$ are i.i.d. with symmetric distribution.

Let $\rho(x)$ be a properly chosen loss function. Then the *M*-estimate of the location parameter μ is the solution of the following optimization problem:

$$\min_{\mu} \sum_{i=1}^{N} \rho(x_i - \mu). \tag{2.1}$$

Defining the derivative $\psi(x) = \rho'(x)$, the corresponding estimating equation is simply

$$\sum_{i=1}^{N} \psi(x_i - \mu) = 0. \tag{2.2}$$

If the density $f(x)$ of the errors is known, then the loss function in (2.1) can be chosen as $\rho(x) = -\ln[f(x)]$, and $\psi(x) = -f'(x)/f(x)$, and the solution to (2.2) is the maximum likelihood estimate (MLE) of the location parameter μ. When the error density is not exactly known, the loss function is usually chosen to be the quadratic $\rho(x) = x^2/2$, and the psi-function is just the identity function. As is well known, the resulting LS-estimate, i.e., the sample mean, is optimal if the error density is Gaussian. The least absolute deviation estimate is obtained by choosing $\rho(x) = |x|$ with $\psi(x) = \text{sgn}(x)$, and the resulting estimate is the sample median, which is optimal if the error distribution is double exponential, or sometimes called Laplacian. The sample mean is very sensitive to the tail behavior of the error distribution, in the sense that a single outlier can alter the estimate arbitrarily. On the other hand, the sample median is sensitive to the behavior of the error distribution at its median.

A more general error criterion is to use the l_p error measure by choosing the loss function $\rho(x) = (1/p)|x|^p$, $1 \le p \le 2$; and

$$\psi_p(x) = |x|^{p-1} \cdot \text{sgn}(x). \tag{2.3}$$

The l_p criterion for small p puts more weights on data that center around the median, however, it still suffers from the lack of robustness against outliers except for the cases where $p = 1$, because $\psi_p(x)$ is bounded only when $p = 1$.

We will now discuss a class of minimax estimators that lie between the sample mean and the sample median. This class of estimators uses the psi-function suggested by Huber [11],

$$\psi_H(x) = \min[c, \max(x, -c)]. \tag{2.4}$$

Huber's psi-function $\psi_H(x)$ has the key robustness properties that it is bounded, monotonely nondecreasing, and continuous. Monotonicity yields uniqueness of *M*-estimate solutions. The effect of using $\psi_H(x)$ is to assign less weight to the small portion of large residuals so that the outliers will not terribly influence the final estimate, while giving unity weight to the bulk of small to moderate residuals. $\psi_H(x)$ is now widely used for obtaining robust M-estimates of location parameter. It will result in an optimal estimate when the error distribution is the least favorable density in Huber's minimax optimization setup [11]

$$f_H(x) = \frac{(1-\gamma)}{\sqrt{2\pi}} \exp^{-\rho_H(x)}. \tag{2.5}$$

It is a mixture distribution such that $(1-\gamma)$ percent of the time the error is from a nominal Gaussian distribution, while γ percent of the time the error is from a distribution G with much larger variance. The result is that it is Gaussian in the middle and Laplacian at the tails, and the associated loss function is

$$\rho_H(x) = \begin{cases} x^2/2 & \text{if } |x| \le c \\ c|x| - c^2/2 & \text{if } |x| > c \end{cases} \tag{2.6}$$

where c is an efficiency *tuning* constant, which is a function of the contaminated percentage γ.

The performance of the Huber's M-estimate is not sensitive to the choice of the tuning constant c in (2.4). For most applications, the contamination percentage and the contaminated distribution are not known. One would like to select the tuning constant to achieve high efficiency both for the nominal Gaussian distribution and for most mixture distributions. This property is called *efficiency robustness*. For Huber's M-estimate, any value between 1.0 and 2.0 will achieve efficiency robustness for $\gamma \le 0.2$. The heavier the tails of the error distribution, the smaller the tuning constant that should be selected. In practice, $\psi_H(x)$ with $c = 1.5$ is often used. Such a choice will give much higher efficiency than the CLP estimate when the underlying distribution is of heavy-tailed non-Gaussian, while it will still maintain well above 96 percent efficiency if the true error distribution is Gaussian.

III. Robust Linear Prediction

Assume that a time series $\{s_1, \cdots, s_N\}$ is generated by the AR(p) model (1.3). A general observation model can be considered

$$y_n = s_n + v_n \tag{3.1}$$

where $\{y_n\}$ is the observation sequence and $\{v_n\}$ is the additive measurement noise sequence. In practice, two distinct outlier models could be formulated, namely, the additive outlier (AO) model and the innovation outlier (IO) model. The AO model is often characterized by patchy outliers due to abrupt change in measurement conditions, or gross error caused by recording instruments. Such outlier bursts usually only affect the observation sequence locally. On the other hand, when $v_n = 0$, we have the IO model where the outlier is generated in the excitation process $\{e_n\}$. Such outliers simply arise from the nature of the physical data generation process, and will affect all the subsequent observations. For voiced speech signal, IO is a more appropriate model, and we will therefore focus our discussion only on the IO model. The reader is referred to Martin [14] and Martin and Yohai [16] for a thorough discussion on AO models and on how to use generalized M-estimate to estimate the parameters of AO models.

In the IO model (3.1), the prediction residuals can be expressed as a function of the LP coefficient vector $\boldsymbol{a}$

$$\epsilon_n(\boldsymbol{a}) = s_n + \sum_{i=1}^{p} a_i s_{n-i} \qquad n = p + 1 \cdots N. \tag{3.2}$$

Now the same formulation for obtaining the M-estimate of location can be generalized to solve for the M-estimate of the LP coefficients. Rather than minimizing the sum of squared residuals as in the CLP, an M-estimate of the LP coefficients is obtained by solving the following optimization problem:

$$\min_{\boldsymbol{a}} \sum_{n=p+1}^{N} \rho(\epsilon_n(\boldsymbol{a})) \tag{3.3}$$

where $\rho(x)$ is an appropriate loss function which is symmetric and has bounded derivative $\psi(x) = \rho'(x)$.

In general, the solution of (3.3) is not scale-invariant in the sense that when the observations s_n's are multiplied by a constant, the resulting estimate is not necessarily equal to the original estimate. If we define $\psi_{\hat{s}}(x) = \hat{s}\psi(x/\hat{s})$, where $\hat{s}$ is a robust scale estimate, we have a scale-invariant version of the estimating equation of the form

$$\sum_{n=p+1}^{N} s_{n-j}\psi_{\hat{s}}(\epsilon_n(\boldsymbol{a})) = 0 \qquad j = 1 \cdots p \tag{3.4}$$

and the solution is called a scale-invariant M-estimate or simply an M-estimate. Scale parameter $\hat{s}$ is usually treated as a nuisance parameter. The reader is referred to Huber [11] for a more detailed discussion on how to properly choose a robust scale estimate.

A. M-Estimation Algorithms

In general, the system of equations (3.4) is nonlinear and iterative methods are required to solve for the coefficient a_i's. Given a preliminary estimate $\tilde{\boldsymbol{a}}$, two possible algorithms are proposed in the following.

1) Newton's Algorithm: Taylor series expansion of $\psi(\epsilon_n(\boldsymbol{a}))$ about the given preliminary estimate $\tilde{\boldsymbol{a}}$ yields

$$\begin{aligned}\psi(\epsilon_n(\boldsymbol{a})) &\approx \psi(\epsilon_n(\tilde{\boldsymbol{a}})) + \Delta\epsilon_n\psi'(\epsilon_n(\tilde{\boldsymbol{a}})) \\ &= \psi(\epsilon_n(\tilde{\boldsymbol{a}})) + \sum_{i=1}^{p} s_{n-i}(a_i - \tilde{a}_i)\,\psi'(\epsilon_n(\tilde{\boldsymbol{a}})).\end{aligned} \tag{3.5}$$

Substituting into estimating equations (3.4), we have a matrix equation

$$C^*\Delta\boldsymbol{a} = -\Delta\boldsymbol{c}^* \tag{3.6}$$

where $\Delta \boldsymbol{a} = \boldsymbol{a} - \tilde{\boldsymbol{a}}$ and the weighted covariance matrix C^* and the correlation vector $\Delta \boldsymbol{c}^*$ are defined as follows:

$$C_{ij}^* = \sum_{n=p+1}^{N} s_{n-i} s_{n-j} \psi'(\epsilon_n(\tilde{\boldsymbol{a}})) \qquad 1 \le i, j \le p \tag{3.7}$$

and

$$\Delta c_j^* = \sum_{n=p+1}^{N} s_{n-j} \psi(\epsilon_n(\tilde{\boldsymbol{a}})) \qquad 1 \le j \le p. \tag{3.8}$$

The approximated Newton's M-estimate can then be obtained by $\hat{\boldsymbol{a}} = \tilde{\boldsymbol{a}} + \Delta \boldsymbol{a} = \tilde{\boldsymbol{a}} - (C^*)^{-1} \Delta \boldsymbol{c}^*$.

2) Iterative Reweighted Least Squares Algorithm (IRLS): In some situations, $\psi'(x)$ is often approximated by a weight function defined as $W(x) = \psi(x)/x$, and (3.4) can then be approximated as a weighted least squares equation

$$\sum_{n=p+1}^{N} s_{n-j} \epsilon_n(\boldsymbol{a}^{(k+1)}) \, W(\epsilon_n(\boldsymbol{a}^{(k)})) = 0 \qquad 1 \le j \le p \tag{3.9}$$

where the residuals are weighted in the estimating equations and the superscript in $\boldsymbol{a}^{(k)}$ indicates the kth iteration of the solution. Convert (3.9) into matrix form, we have

$$C^{**} \boldsymbol{a}^{(k+1)} = -\boldsymbol{c}^{**} \tag{3.10}$$

where the weighted covariance matrix C^{**} and the correlation vector $\boldsymbol{c}^{**}$ are defined as

$$C_{ij}^{**} = \sum_{n=p+1}^{N} s_{n-i} s_{n-j} W(\epsilon_n(\boldsymbol{a}^{(k)})) \qquad 1 \le i, j \le p \tag{3.11}$$

and

$$c_j^{**} = \sum_{n=p+1}^{N} s_{n-j} s_n W(\epsilon_n(\boldsymbol{a}^{(k)})) \qquad 1 \le j \le p. \tag{3.12}$$

The IRLS solution is simply $\hat{\boldsymbol{a}}^{(k+1)} = -(C^{**})^{-1} \boldsymbol{c}^{**}$.

B. *Efficiency Robustness of M-Estimation*

When $\rho(t) = t^2/2$, and the associated $\psi(t) = t$, (3.3) and (3.4) yield an LS-estimate $\hat{\boldsymbol{a}}_{LS}$ of $\boldsymbol{a}$, which is the solution of the covariance method. It is well known that under mild regularity assumptions concerning excitation $\{e_n\}$, $\hat{\boldsymbol{a}}_{LS}$ is consistent and asymptotically normal [17]. Furthermore, the asymptotic covariance matrix $V_{LS}(\boldsymbol{a})$ for $\hat{\boldsymbol{a}}_{LS}$ depends only on $\boldsymbol{a}$, and not on the distribution of the innovations [18]. This distribution-free property of $\hat{\boldsymbol{a}}_{LS}$ is very attractive, on one hand, but the LS-estimate nonetheless can be quite inefficient, on the other hand. The following observations reveal the lack of efficiency robustness of $\hat{\boldsymbol{a}}_{LS}$. The asymptotic *Cramer-Rao lower bound* matrix $V_{CR}(\boldsymbol{a})$ for $\boldsymbol{a}$ can be derived as [19], [16]

$$V_{CR}(\boldsymbol{a}) = V_{LS}(\boldsymbol{a}) / [i(F) * \sigma_\epsilon^2] \tag{3.13}$$

with $i(F)$ being the *Fisher information* of the source distribution F, defined as

$$i(F) = E_F[-f'(t)/f(t)]^2. \tag{3.14}$$

Let $V(\hat{\boldsymbol{a}})$ denote the asymptotic covariance matrix of any estimate $\hat{\boldsymbol{a}}$, and adopt

$$AEFF(\hat{\boldsymbol{a}}, F) = \left\{ \det\left[V_{CR}(\boldsymbol{a})\right] / \det\left[V(\hat{\boldsymbol{a}})\right] \right\}^{1/p} \tag{3.15}$$

as the definition of the asymptotic efficiency. Then we have

$$AEFF(\hat{\boldsymbol{a}}_{LS}, F) = \frac{1}{[i(F) * \sigma_\epsilon^2]}. \tag{3.16}$$

This is the same as the asymptotic efficiency of the sample mean in the i.i.d. case, which can be arbitrarily close to zero for finite variance innovations (and equal to zero for infinite variance distributions).

On the other hand, the asymptotic covariance matrix for the M-estimate defined in (3.4) can be derived as [14] and [15]

$$V_M(\hat{\boldsymbol{a}}) = V_{LS}(\hat{\boldsymbol{a}}) * V_{loc}(\psi_s, F) / \sigma_\epsilon^2 \tag{3.17}$$

with $V_{loc}(\psi_s, F)$ being the asymptotic variance for the M-estimate of the location parameter defined as [11]

$$V_{loc}(\psi_s, F) = E_F[\psi_s^2(t)] / E_F^2[\psi_s'(t)]. \tag{3.18}$$

Therefore, we have the asymptotic efficiency on an M-estimate for $\boldsymbol{a}$ as

$$AEFF(\hat{\boldsymbol{a}}_M, F) = \frac{1}{[i(F) * V_{loc}(\psi_s, F)]} \tag{3.19}$$

which is the same as that of an M-estimate for the location parameter. To compare performance, we also define the asymptotic relative efficiency of the M-estimate versus the LS-estimate as the ratio of their asymptotic efficiencies, i.e.,

$$AREFF(\hat{\boldsymbol{a}}_M, \hat{\boldsymbol{a}}_{LS}, F) = \frac{\sigma_\epsilon^2}{V_{loc}(\psi_s, F)}. \tag{3.20}$$

It is noted that for bounded $\psi(x)$, $V_{loc}(\psi_s, F)$ is usually smaller than the innovation variance σ_ϵ^2 for heavy-tailed non-Gaussian distributions. Therefore, the M-estimate is usually more efficient than the LS-estimate when the innovation distribution is of heavy-tailed non-Gaussian and sometimes much more efficient.

Table I lists the Fisher information numbers, the asymptotic efficiencies for the LS-estimate, and Huber's M-estimate (with $c = 1.5$), and their asymptotic relative efficiencies for a number of γ-contaminated normal distributions $CN(\gamma, \sigma)$. The values of $i(F)$ and $AEFF(\hat{\boldsymbol{a}}_{MH}, F)$ are very stable, while $AEFF(\hat{\boldsymbol{a}}_{LS}, F)$ varies drastically. For cases with big contaminated variance and/or moderate contaminated percentage, the LS-estimate $\hat{\boldsymbol{a}}_{LS}$

TABLE I
ASYMPTOTIC EFFICIENCY COMPARISONS FOR $CN(\gamma, \sigma)$

γ	σ	$i(F)$	$AEFF(\hat{a}_{LS})$	$AEFF(\hat{a}_{MH})$	$AREFF$
0.0	—	1.0	1.0	0.964	0.964
0.1	3	0.796	0.698	0.968	1.387
0.1	6	0.805	0.276	0.875	3.170
0.1	10	0.827	0.111	0.819	7.379
0.25	3	0.590	0.565	0.906	1.603
0.25	6	0.593	0.173	0.680	3.931
0.25	10	0.625	0.062	0.567	9.145

can be terribly inefficient. For instance, with $(\gamma, \sigma) = (0.1, 10)$, we have $i(F) = 0.827$ and $\sigma_\epsilon^2 = 10.9$, and therefore, the asymptotic efficiency for $\hat{a}_{LS}$ is only about 11 percent. By using Huber's $\psi_H(x)$ with the tuning constant $c = 1.5$, the asymptotic efficiency is 82 percent. Compared to the LS-estimate $\hat{a}_{LS}$, the Huber's M-estimate $\hat{a}_{MH}$ is about 738 percent more efficient.

C. *Bias Reduction of M-Estimation*

When the excitation sequence $\{e_n\}$ is quasi-periodic, as in the cases for modeling voiced speech signal, this periodic innovation sequence will often couple with the linear filter and can result in "interference" across the pitch periods which in term results in bias in the LPC estimates. By replacing the prediction residual in (3.9) with (3.2), the bias vector can be solved with the following matrix equation:

$$C^{**}\Delta \boldsymbol{a} = -\Delta \boldsymbol{c}^{**} \tag{3.21}$$

where $\Delta \boldsymbol{a}$ is the bias vector, and

$$\Delta c_j^{**} = \sum_{n=p+1}^{N} s_{n-j} e_n W(\epsilon_n(\tilde{\boldsymbol{a}})) \qquad 1 \le j \le p. \tag{3.22}$$

The right-hand side of (3.22) is a weighted correlation between the unobservable excitation sequence $\{e_n\}$ and the observation sequence $\{s_{n-j}\}$. For white noise excitations, this correlation is asymptotically zero for both the LS- and M-estimates. However, for periodic excitations, the correlation term is no longer zero for high-pitched voices due to more often constructive interference from the periodic components. The M-estimation procedure generally assigns less weight to larger excitations and therefore results in less bias than the conventional LP procedure where the weights are assigned equally. An extreme example is to set zero weight as in the sample-selective linear prediction [20], when the excitation exceeds some threshold; and, in these cases, the bias can be shown to be negligible. Numerical simulation indicates that even for high-pitched voices, the bias could be reduced effectively using M-estimate. Some numerical simulation examples can be found in Section V.

IV. ALGORITHMIC ISSUES FOR SPEECH ANALYSIS

The RBLP algorithm discussed in the previous section is rather general. For speech signal analysis, one is more concerned with certain specific issues. For instance, computation complexity is an important topic for real-time speech analysis systems. Stability of the all-pole filter is of great concern for speech coding. The validity of the AR(p) assumption for nonstationary speech signal could also pose a problem in analyzing the transient part of the speech signal. We will discuss those algorithmic issues in the following sections.

A. *Model Validity*

From our discussion in Section III, the RBLP algorithm offers a much better fit for steady-state vowels, especially for the cases with a small pitch period and/or with high peak factor. For other steady-state sounds where the source excitation can be well approximated by Gaussian innovations, the RBLP will perform nearly as well as the CLP procedures. For the transient part of the speech signal, both the robust and the conventional LP procedures give only an approximation to the nonstationary model.

B. *Stability*

One of the greatest concerns for speech synthesis is the stability of the linear prediction synthesis filter. Only the conventional autocorrelation method and lattice methods guarantee stability. However, stability checks used in conventional covariance methods can also be incorporated into the RBLP procedures. If the RBLP algorithm produces an unstable LP filter, then the procedure can be stopped, and the stable preliminary LP filter is then used in the synthesis filter.

C. *Computation Complexity*

Except for the estimation of $\hat{s}$, the computation of the prediction residuals $\epsilon_n(\boldsymbol{a})$, and the evaluation of $\psi(x)$ and $\psi'(x)$ at all $(N - p)$ residuals, the computation complexity of one iteration of Newton's method and IRLS method are the same as that of the covariance method. Fast algorithms for solving covariance equations are readily available. The prediction residual can be computed by inverse filtering. The scale can be chosen as the standard deviation of the weighted residuals or any robust scale estimate suggested in [11]. And, by choosing Huber's psi-function (2.4), the evaluation of $\psi(x)$ and $\psi'(x)$ can be greatly simplified. If the l_p error criterion is used, then fast l_p deconvolution algorithms are also available [21].

D. *Approximated M-Estimation for the LP Coefficients*

Since $\epsilon_n(\boldsymbol{a})$ is asymptotically uncorrelated with s_{n-i}, the weighted term $\psi'(\epsilon_n(\boldsymbol{a}))$ can be replaced by an estimate of the weighted without affecting the solution too much. If one uses

$$A^*(\boldsymbol{a}) = \frac{1}{(N-p)} \sum_{n=p+1}^{N} \psi'(\epsilon_n(\boldsymbol{a})), \tag{4.1}$$

(3.5) can now be simplified as

$$C\Delta \boldsymbol{a} = -\left[A^*(\tilde{\boldsymbol{a}})\right]^{-1} \Delta \boldsymbol{c}^* \tag{4.2}$$

where C is simply the covariance matrix in the covariance method. The procedure to obtain $\hat{\boldsymbol{a}}$ through (4.2) is called an *approximated Newton's method*. It is noted that C^{-1} need not be recomputed at all subsequent M-estimation iterations in order to solve (4.2), therefore, the computational complexity can be greatly reduced.

IRLS can also be approximated the same way as Newton's method to give an approximated IRLS estimate $\hat{\boldsymbol{a}}$ such that

$$C\hat{\boldsymbol{a}} = -\left[A^{**}(\tilde{\boldsymbol{a}})\right]^{-1}\boldsymbol{c}^{**} \tag{4.3}$$

with the modified weight

$$A^{**}(\boldsymbol{a}) = \frac{1}{(N-p)} \sum_{n=p+1}^{N} \psi(\epsilon_n(\boldsymbol{a}))/\epsilon_n(\boldsymbol{a}). \tag{4.4}$$

In the case of speech analysis, the preliminary $\tilde{\boldsymbol{a}}$ can simply be chosen as the solution of the CLP procedure. A one-step M-estimate with a single iteration of (4.2) or (4.3) usually provides an adequate solution.

E. Special Cases of Robust Linear Prediction for Speech Analysis

Some special forms of robust M-estimates for the LP coefficients have been suggested in the literature. In [20], a sample-selective linear prediction (SSLP) algorithm was used to improve the source and transfer function separation so that better bandwidth, and hence better vocal tract area function, could be obtained. The SSLP algorithm is a special form of the RBLP using the hard rejection psi-function.

$$\psi_{HR}(x) = \begin{cases} x & \text{if } x \leq c \\ 0 & \text{if } x > c. \end{cases} \tag{4.5}$$

For synthetic data, the SSLP usually performs very well. However, for natural speech, the choice of the rejection points is very crucial. Multiple roots will become a serious problem if preliminary solution is not chosen carefully. Stability is also of great concern. Kang and Everett [22] realized that the inclusion of the intraframe pitch interference (outliers) in the CLP analysis leads to broadened resonant bandwidths and makes the synthesized speech sound fuzzy. They then used the SSLP on natural speech and claimed to produce more "focused" speech quality. The SSLP was also shown to be effective for high-pitched voices, and produced more compact vowel clusters in the $F_1 - F_2$ plane [23]. In [24], a more general version of SSLP, called weighted linear prediction (WLP), is proposed, in which the weight is a continuous function of the prediction residual. The WLP algorithm uses an ad hoc method to select the weights in (3.11) and (3.12) in order to solve the IRLS algorithm (3.10), and can be considered as a special case of RBLP. The WLP performs quite well for both formant estimation of synthetic vowels and for vowel recognition of natural human speech [24]. Finally, a least absolute error criterion, which is a special form of the RBLP with psi-function chosen as the SGN function, has also been applied to linear prediction [25]. The result is that the residuals show more peaky characteristics, which can be used to detect pitch effectively.

V. Performance Evaluation

The RBLP algorithm proposed has been tested on both synthetic and natural speech. For purposes of comparison, two known sets of synthetic data were used to verify the effectiveness of the proposed RBLP procedure.

A. Testing Results on Klatt's Synthetic Data

It is well known that the formant frequencies and bandwidths are not easy to measure as fundamental frequency (F_0) varies in speech. To compare the performance of three known speech analysis algorithms, namely, HAMON, filter bank, and LPC analysis, Klatt [26] uses a set of fixed synthesis parameters to generate typical synthetic vowel sounds in the word "bit." The formant center frequencies F_i and bandwidths B_i used are $F_1 = 400$ Hz, $B_1 = 50$ Hz, $F_2 = 1800$ Hz, $B_2 = 140$ Hz, $F_3 = 2900$ Hz, $B_3 = 240$ Hz, $F_4 = 3800$ Hz, and $B_4 = 350$ Hz, with F_0 varying in steps from 133 Hz to 200 Hz. Klatt's conclusion was that for all the three algorithms compared, there is a tendency for the first formant to be biased toward the frequency of the most intense harmonic, resulting in an error of up to ±8 percent for this excitation set.

In this paper, we are more interested in comparing the performance of the CLP and the RBLP procedures. In addition to the nine F_0's (200, 189, 179, 169, 160, 152, 145, 139, 133 Hz) used in [26], eight more F_0's ranging from 225 Hz to 400 Hz in increments of 25 Hz are also used to generate the testing synthetic vowels. The sampling frequency is 10 kHz, the data window used is a 25.6 ms Hamming window, and the analysis procedure is a 14-pole conventional autocorrelation method. In the following, the error is defined as $E = |\hat{x} - x_0|/x_0$, where $\hat{x}$ is the estimated parameter value and x_0 is the true parameter value. The results of the CLP shows that the F_1 errors range from -12 to $+9$ percent, with an average absolute error of 4.7 percent. As for the RBLP analysis, we use the data windowed by the same Hamming window, and Huber's psi-function with $c = 1.5$ to extract the M-estimate for the LP coefficients. For a one-step M-estimate, i.e., one M-estimation iteration of (3.6), the F_1 errors are within ±3 Hz, with an average absolute error rate of 0.4 percent. In Table II, we compare the results of estimating the first formant frequency and bandwidth and list the maximum and average absolute error rates over all F_0 values. The second column of Table II summarizes the CLP results, and the third and fourth columns correspond to the one-step and the 6-step M-estimates. There is clearly a tremendous improvement over the CLP results. The maximum F_1 bias is reduced from 48 Hz to 4 Hz in just one M-estimation iteration, which demonstrates that M-estimate reduces bias effectively. The linear prediction spectra for both the CLP estimate and the one-step M-estimate are plotted in Fig. 1(a) and (b). For both plots, the 17 curves correspond to the 17 F_0 values tested, with F_0 decreasing from left to right. The leftmost and rightmost curves are associated with $F_0 = 400$ Hz and 133 Hz, respectively. The top plot shows that estimated F_1 fluc-

TABLE II
SUMMARY OF THE CLP AND THE RBLP ON KLATT'S DATA

Error Rate	iter = 0	iter = 1	iter = 6
MAX $\|\Delta F_1\|$ (Hz)	48	4	1
MAX $\|\Delta B_1\|$ (Hz)	170	18	8
$\|\Delta F_1\|/F_1 * 100$ percent	4.7	0.39	0.08
$\|\Delta B_1\|/B_1 * 100$ percent	97.0	10.6	7.9

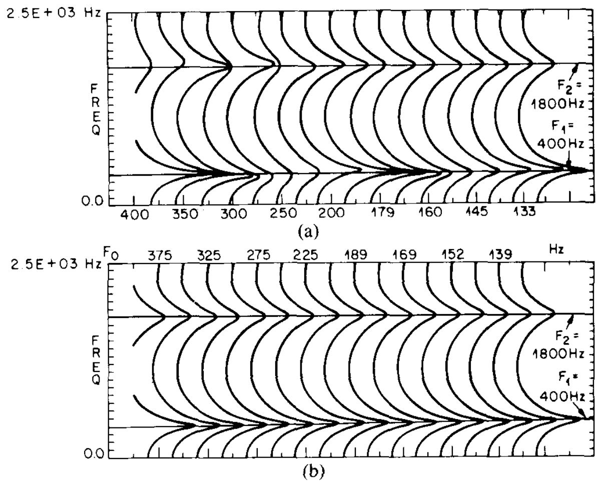


Fig. 1. The LP spectrum plots for Klatt's synthetic vowels. (a) Top—CLP analysis. (b) Bottom—RBLP analysis.

tuates around 400 Hz, with accurate F_1 estimation at F_0 = 400, 200, and 133 Hz, where F_1 is a multiple of F_0. The estimated F_2 error is usually smaller than the F_1 error. It also illustrates the fluctuation of the estimated B_1 values, and that there is a strong tendency to underestimate B_1 when F_1 is a multiple of F_0, and to overestimate B_1 when $F_1 = (k + \frac{1}{2})F_0$ [27]. In the bottom plot, the spectra of the one-step M-estimates show consistent estimation over the wide range of F_0 values tested.

It is noted that windowing destroys the original structure of the AR(p) process in the sense that the windowed data is no longer an AR(p) process with the original model parameters. However, the RBLP procedure still gives a fairly good approximation to the original AR(p) spectrum of the vocal tract filter. This suggests that the RBLP procedure can be applied to real speech data even when the autoregressive model is only an approximation to the speech signal.

B. Testing Results on Atal's Synthetic Data

In [27], Atal investigated the influence of pitch periodicity on the estimation of F_1 and B_1. The data used were generated by exciting a 2-pole filter with a periodic pulse train. The center frequency varied from 200 Hz to 700 Hz, while the bandwidth was fixed at 50 Hz. Three pitch values, F_0 = 400, 200, and 100 Hz, were used. By using a pitch synchronous covariance method performed on data spanning exactly one pitch period, it was shown that the estimated F_1 error can be greater than 10 percent for high-pitch values, i.e., $F_0 \geq 200$ Hz, which is more than 3–5 percent JND (just noticeable difference) reported by Flanagan [4]. The estimated bandwidth exhibits much more variation due to the influence of the periodic excitation. However, if the analysis is performed by a window that excludes the sample at the excitation, the formant and bandwidth estimation errors can be shown to be negligibly small at all pitch frequencies. It is noted that a single spiky excitation is enough to make the estimated formant deviate by more than 10 percent from the correct solution. If pitch-asynchronous analysis is desired, then for high-pitched voiced sounds, there is usually more than one spiky excitation in a fixed-length analysis window. Under these conditions, the CLP procedure will generally give significant errors in formant and bandwidth estimates.

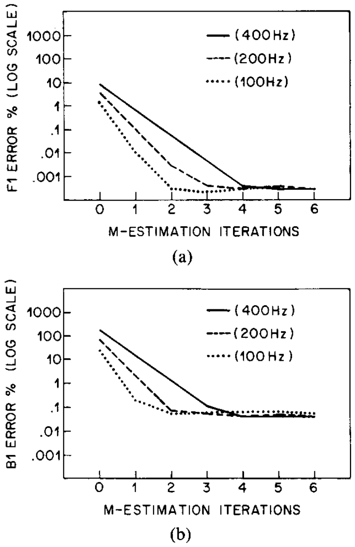
The same setup was used to compare the conventional covariance method to the RBLP procedure. Pitch asynchronous analysis was performed in all cases. The sampling frequency was 10 kHz, the analysis frame was again 25.6 ms, and the analysis filter order was fixed at 10. For the robust M-estimate, we still use Huber's psi-function with c = 1.5. Table III summarizes the maximum and average absolute error rates of F_1 and B_1 over all F_1 values tested. It indicates that the one-step M-estimate generally reduces the maximum bias and the average error rate by more than an order of magnitude. The convergence of the M-estimate is shown in Fig. 2, where we plot the logarithm of the average formant and bandwidth error rates versus the number of iterations for all three F_0 values. The zeroth iteration corresponds to results from the CLP procedure. After four M-estimation iterations, the estimated errors converge to the same values for all three pitch frequencies tested. This confirms our theoretical analysis that the RBLP procedure is not sensitive to the location of the pitch excitation and the value of the pitch frequency.

C. Testing on Natural Human Speech

The CLP procedures are known to be sensitive to the location of the analysis window and the value of the pitch frequency. However, from the discussion in the previous sections, we expect the RBLP procedure to be less affected by the placement of the window and the number of pitch periods included in the analysis window. As an illustration, line spectrum pairs (LSP) [28] obtained from both procedures are plotted in Fig. 3. The utterance is the word "one" spoken by a female speaker with pitch frequency around 180 Hz. The sampling rate is 10 kHz, and 12-order LPC analysis without preemphasis is performed every 1 ms on a frame of 20 ms. The top LSP plot is obtained from the conventional autocorrelation method using a 20 ms Hamming window, and the bottom plot is obtained from an M-estimate using Huber's psi-function with c = 1.5 without applying any window. It is clearly shown in the plots that the LSP's obtained from the RBLP algorithm exhibit less variability over time, while CLP analysis gives more local variations due to the positioning of the window. It is expected that higher coding efficiency can be achieved by using the robust LSP's.

TABLE III
SUMMARY OF THE CLP AND THE RBLP ON ATAL'S DATA

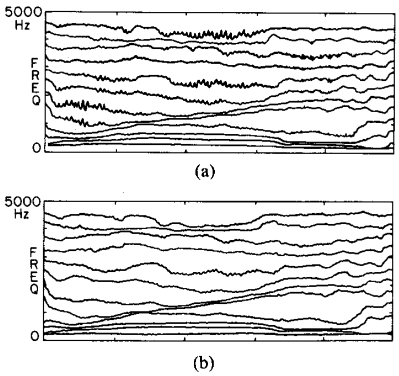
	Conventional LP			One-Step M-Estimate		
Pitch F_0 (Hz)	400	200	100	400	200	100
MAX $\lvert \Delta F_1 \rvert$ (Hz)	56	25	12	5	1	0.1
MAX $\lvert \Delta B_1 \rvert$ (Hz)	190	67	24	16	2	0.3
$\lvert \Delta F_1 \rvert / F_1$ (percent)	9	4	1	1	0.1	0.01
$\lvert \Delta B_1 \rvert / B_1$ (percent)	189	70	24	15	2	0.2



(a)

(b)

Fig. 2. The average error rates versus M-estimation iterations. (a) Top—formant error. (b) Bottom—bandwidth error.



(a)

(b)

Fig. 3. The LSP plots of "one" uttered by a female speaker. (a) Top—CLP analysis. (b) Bottom—RBLP analysis.

VI. Conclusion

We have shown that the RBLP algorithm produces a less biased estimate of the linear prediction coefficients than the CLP algorithm. Formant frequency and bandwidth estimation based on solving the roots of the robust prediction polynomial will therefore give more accurate results. We have also demonstrated that better source excitation and vocal tract filter separation can be achieved by using the RBLP procedures and the influence of voice periodicity on the formant and bandwidth estimation is also reduced. The RBLP algorithm proposed is also less sensitive to the selection of analysis window, the location of the excitation, and the value of the pitch frequency. Testing on synthetic vowel data indicates that the RBLP procedure is able to reduce the formant frequency and bandwidth error rates by more than an order of magnitude. More experiments are required to evaluate the performance of the RBLP algorithm for natural human speech analysis.

Acknowledgment

The author would like to thank Prof. R. D. Martin of the University of Washington for stimulating discussion throughout this work. The author is also indebted to his colleague Dr. F. K. Soong for preparing Fig. 3 and sharing some of his plotting software.

References

[1] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," in *Rep. 6th Int. Congr. Acoust.*, Y. Kohasi, Ed., Tokyo, Japan, C-5-5, C17-20, 1968. (Also see *Speech Synthesis*, J. L. Flanagan and L. R. Rabiner, Eds. Stroudsburg, PA: Dowden, Hutchinson, and Ross, 1973, pp. 289-292.)

[2] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc. 1967 Conf. Commun. and Processing*, 1967, pp. 360-361.

[3] G. C. M. Fant, *Acoustic Theory of Speech Production*. Gravenhage, The Netherlands: Mouton, 1960.

[4] J. L. Flanagan, *Speech Analysis, Synthesis, and Perceptions*, 2nd ed. New York: Springer-Verlag, 1972.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, pp. 561-580, 1975.

[6] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[7] J. W. Tukey, "A survey of sampling from contaminated distributions," in *Contributions to Probability and Statistics*, X. Olkin, Ed. Stanford, CA: Stanford University Press, 1960, pp. 448-484.

[8] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Tech. J.*, vol. 64, no. 4, pp. 1211-1234, 1985.

[9] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 4, pp. 1235-1249, 1985.

[10] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[11] ——, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73-101, 1964.

[12] ——, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753-1758, 1965.

[13] ——, "Robust regression: Asymptotics, conjectures and Monte Carlo," *Ann. Statist.*, vol. 1, pp. 799-821, 1973.

[14] R. D. Martin, "Robust methods for time series," in *Applied Time Series II*, D. F. Findley, Ed. New York: Academic, 1981, pp. 683-759.

[15] R. D. Martin and V. J. Yohai, "Robustness in time series and estimating ARMA models," in *Handbook of Statistics*, vol. 5, E. J. Hannan, P. R. Krishnaiah, and M. M. Rao, Eds. Amsterdam, The Netherlands: Elsevier Science, 1985, pp. 119-155.

[16] C.-H. Lee, "M-estimate for ARMA process," Ph.D. dissertation, Dep. Elec. Eng., Univ. Washington, Seattle, 1981.

[17] T. W. Anderson, *The Statistical Analysis of Time Series*. New York: Wiley, 1971.

[18] P. Whittle, "Gaussian estimation in stationary time series," *Bull Int. Stat. Inst.*, vol. 39, pp. 105-129, 1962.

[19] R. D. Martin, "The Cramer-Rao bound and robust *M*-estimates for autoregressions," *Biometrika*, vol. 6, no. 2, pp. 437-442, 1982.

[20] R. Mizoguchi, M. Yanagida, and O. Kakusho, "Speech analysis by selective linear prediction in the time domain," in *Proc. ICASSP'82*, Paris, France, 1982, pp. 1573-1576.

[21] R. Yarlagadda, J. Bee Bednar, and T. L. Watt, "Fast algorithms for l_p deconvolution," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 174-182, Feb. 1985.

[22] G. S. Kang and S. S. Everett, "Improvement of the LPC analysis," in *Proc. ICASSP'83*, Boston, MA, 1983, pp. 89-92.

[23] Y. Miyoshi *et al.*, "Analysis of speech signal of short pitch period by the sample selective linear prediction," in *Proc. ICASSP'86*, Tokyo, Japan, 1986, pp. 1245-1248.

[24] M. Yanagida and O. Kakusho, "A weighted linear prediction analysis of speech signals by using the Given's reduction," presented at the IASTED Int. Symp. Appl. Signal Processing and Digital Filtering, Paris, France, June 19-21, 1985.

[25] E. Denoel and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1397-1403, Dec. 1985.

[26] D. H. Klatt, "Representation of the first formant in speech recognition and in models of the auditory periphery," in *Proc. Montreal Symp. on Speech Recogn.*, 1986, pp. 3-5.

[27] B. S. Atal, "Linear prediction of speech—Recent advances with applications to speech analysis," in *Speech Recognition*, D. R. Reddy, Ed. New York: Academic, 1975, pp. 221-230.

[28] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP'84*, San Diego, CA, 1984, pp. 1.10.1-4.

Chin-Hui Lee (S'79-M'81) was born in July 1951. He received the B.S. degree from National Taiwan University, Taipei, in 1973, the M.S. degree from Yale University, New Haven, CT, in 1977, and the Ph.D. degree from the University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981 he joined Verbex Corporation, Bedford, MA, and was involved in research work on connected word recognition. In 1984 he became affiliated with Digital Sound Corporation, Santa Barbara, CA, where he engaged in research work in speech coding, speech recognition, and signal processing for the development of the DSC-2000 Voice Server. Since 1986 he has been with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech modeling, speech recognition, and signal processing.