A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification

Ravi P. Ramachandran, Mihailo S. Zilovic, and Richard J. Mammone, *Senior Member, IEEE*

***Abstract*—In this paper, various linear predictive (LP) analysis methods are studied and compared from the points of view of robustness to noise and of application to speaker identification. The key to the success of LP techniques is in separating the vocal tract information from the pitch information present in a speech signal even under noisy conditions. In addition to considering the conventional, one-shot weighted least-squares methods, we propose three other approaches with the above point as a motivation. The first is an iterative approach that leads to the weighted least absolute value solution. The second is an extension of the one-shot least-squares approach and achieves an iterative update of the weights. The update is a function of the residual and is based on minimizing a Mahalanobis distance. Third, the weighted total least-squares formulation is considered. A study of the deviations in the LP parameters is done when noise (white Gaussian and impulsive) is added to the speech. It is revealed that the most robust method depends on the type of noise. Closed-set speaker identification experiments with 20 speakers are conducted using a vector quantizer classifier trained on clean speech. The relative performance of the various LP approaches depends on the type of speech material used for testing.**

I. Introduction

A LINEAR predictive (LP) analysis [1] of a speech signal is based on the all-pole model. This model assumes that a speech sample is a weighted linear combination of p previous samples and is described by the equation

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \tag{1}$$

where $s(n)$ is the speech signal, $e(n)$ is the error or LP residual, and a_i are the weights applied to the previous speech samples. The weights correspond to the direct-form coefficients of a nonrecursive filter $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$. Passing the speech signal through the filter $A(z)$ results in the removal of the near-sample correlations and produces the LP residual $e(n)$. The LP residual represents the pitch information in the speech. The magnitude spectrum of $1/A(z)$ describes the spectral envelope of the speech. In particular, this gives information about the formant frequencies that characterize the vocal tract resonances. Therefore, LP analysis has both time- and frequency-domain interpretations.

Manuscript received November 28, 1993; revised July 11, 1994. This work was supported by the Air Force/Rome Laboratories.

The authors are with CAIP Center, Department of Electrical Engineering, Rutgers University, Piscataway, NJ 0885-1390 USA.

IEEE Log Number 9408402.

Since LP analysis leads to a description of the spectral envelope, it finds application in different aspects of speech processing. These applications include predictive speech coding [2], speech enhancement [3], and speaker recognition [4], [5]. The success of LP methods is in determining the coefficients a_i such that 1) $A(z)$ captures the vocal tract information and 2) the LP residual contains the pitch information. This is equivalent to achieving a clear separation of the vocal tract information and the pitch information, both of which are contained in the speech signal. In addition, LP methods must be robust to noise in that the vocal tract information should be extracted even for noisy speech. It has been observed that the conventional method of LP analysis based on squared error is sensitive to noisy speech [6].

This paper addresses the issue of robust LP methods by comparing four different approaches. First, we consider the conventional approach of minimizing the weighted mean-square LP residual. Second, we use a method based on minimizing the weighted absolute value of the LP residual. Third, we study the formulation of an iterative method based on minimizing the weighted mean-square LP residual. The weights are iteratively updated as the inverse covariance matrix of the LP residual. Fourth, we apply the weighted total least squares approach that still produces an all-pole LP filter $1/A(z)$ but assumes that the error not only occurs in the given observation (the speech sample) but also in the data (the previous samples). The theoretical formulation of each method, the implementational aspects, and performance comparisons are discussed. From the point of view of application, we focus on the features applied to speaker recognition in that the LP parameters are used to identify a speaker from his or her voice. A robust parameter set allows for the successful recognition of a speaker even if the speech is noisy.

The outline of this paper is as follows. Section II discusses the theoretical aspects of the various LP methods considered. In Section III, the experimental results are provided with respect to additive noise effects. The results of speaker identification experiments are discussed in Section IV. Section V provides the summary and conclusions.

II. Linear Predictive Analysis Methods

The purpose of this section is to describe the various LP analysis methods we consider. The theoretical formulation,

the method of solution, and some implementational aspects are discussed.

A. Conventional LP Solution Based on Weighted Squared Error

The most well-known technique in LP estimation is based on minimizing the weighted mean-squared error, which is expressed as

$$E_2 = \sum_{n=n_1}^{n_2} w(n)e^2(n) = \sum_{n=n_1}^{n_2} w(n)\,(s(n) - \boldsymbol{S}^T\boldsymbol{a})^2 \qquad (2)$$

where $s(n)$ is the nth sample of the speech signal, $\boldsymbol{a} = [a_1 a_2 \cdots a_p]^T$ is the vector of the predictor coefficients, $\boldsymbol{S} = [s(n-1)s(n-2)\cdots s(n-p)]^T$, and p denotes the order of the prediction. The limits n_1 and n_2 are the starting and ending time indices of a segment or frame of the speech. The coefficients a_i for $1 \le i \le p$ can be obtained by finding $\partial E_2/\partial a_i$ and equating it to zero. This leads to the set of p equations of the form

$$\sum_{n=n_1}^{n_2} w(n)s(n-i)\boldsymbol{S}^T\boldsymbol{a} = \sum_{n=n_1}^{n_2} w(n)s(n-i)s(n) \qquad i \in (1, 2, \cdots, p). \qquad (3)$$

This set of equations can be put in a vector form as

$$\sum_{n=n_1}^{n_2} w(n)\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{a} = \sum_{n=n_1}^{n_2} w(n)\boldsymbol{S}s(n). \qquad (4)$$

This gives a closed-form solution for the predictor coefficients [7]:

$$a = \left(\sum_{n=n_1}^{n_2} w(n)\boldsymbol{S}\boldsymbol{S}^T\right)^{-1}\left(\sum_{n=n_1}^{n_2} w(n)\boldsymbol{S}s(n)\right). \qquad (5)$$

Based on the limits n_1 and n_2 of the summations in (3)–(5), two different cases can be analyzed. The autocorrelation method results when $n_1 = 1$ and $n_2 = N + p$, where N is the length of the analysis frame. The signal is assumed to be zero outside of the interval $[1, N]$. The covariance method results when $n_1 = 1$ and $n_2 = N$. Therefore, in the autocorrelation case, the first summation on the right-hand side of (5) represents a positive-definite, symmetric Toeplitz matrix whereas in the covariance case, it represents a positive semi-definite, symmetric matrix [8]. There is little computational effort in determining $\boldsymbol{a}$ in that for the autocorrelation case, an efficient Levinson–Durbin recursion is used, and for the covariance case, the Cholesky decomposition is used to solve the system of equations [1]. In implementing both the autocorrelation and covariance approaches, we set $w(n)$ to be the Hamming window weights. In the sequel, we refer to the autocorrelation approach as the weighted mean-squared error autocorrelation (WMSEA) method. A similar acronym for the covariance method is WMSEC.

B. Weighted Least Absolute Value Minimization

We consider the use of a weighted L_1 norm of the error in which the objective function to be minimized is

$$E_1 = \sum_{n=n_1}^{n_2} w(n)|e(n)| = \sum_{n=n_1}^{n_2} w(n)|s(n) - \boldsymbol{S}^T\boldsymbol{a}|. \qquad (6)$$

For computation of the coefficients a_i, the problem is set up as a linear program in that the function E_1 is minimized subject to the constraint of (1) [8]. The linear program is solved by the simplex method [9], which examines different feasible solutions and picks the one that leads to the smallest value of E_1. Note that although one solution is picked, it need not be unique. The simplex method does not necessarily examine all feasible solutions but moves from one solution to another along a connecting edge of a polyhedra to decrease E_1 as much as possible [9]. This procedure makes the process iterative and hence is computationally more expensive than minimizing the weighted mean-squared error. For our experiments, we use the covariance type formulation ($n_1 = 1$ and $n_2 = N$) with Hamming window weights $w(n)$.

The motivation of considering an L_1 type objective function is that it attaches less importance to the outliers than its L_2 counterpart described above. This will allow for the pitch pulses in the original speech signal to be relatively ignored for the prediction and hence put more pitch information in the LP residual. Similarly, bursts of noise in the speech will be more reflected in the residual, thereby preserving the vocal tract information in the parameter set. Minimizing an L_1 type objective function has been used in the context of restoring the spectrum of a diffraction-limited image [10], image restoration [8], improving the frequency resolution of spectral estimates [11] and in determining the multipulse excitation for a predictive speech coder [12]. In [13], linear predictive spectral estimation using the L_1 norm is used to separate two sinusoids in noise. Results show that when impulsive noise is present, an L_1 estimator can resolve the sinusoids, whereas an L_2 estimator cannot. This method is referred to as the weighted least absolute value (WLAV) method.

C. Iterative Weighted Least-Squares Approach

A natural extension to the conventional weighted mean-squared error solution is to achieve an iterative update of the weights $w(n)$ until a steady set of predictor coefficients is found. This is referred to as the iterative weighted least-squares (IWLS) approach. The steps are as follows:

- Step 1: For the first iteration, compute $\boldsymbol{a}$ by the standard WMSEC approach using unity weights (see (5)).
- Step 2: Given $\boldsymbol{a}$, determine the error signal or LP residual $e(n)$ using (1).
- Step 3: Compute the diagonal elements of the covariance matrix corresponding to the LP residual.
- Step 4: If the eigenvalue spread is greater than 100, readjust the spread to be 100 by upward scaling of the smaller eigenvalues.

- Step 5: For the third and subsequent iterations, smooth the covariance matrix by taking the matrix from the previous iteration into account.
- Step 6: Set the weights $w(n)$ to be the reciprocal of the diagonal elements of the smoothed covariance matrix. Recompute $\boldsymbol{a}$ by the standard WMSEC method using these new weights.
- Step 7: If the norm of the difference between the current estimate of $\boldsymbol{a}$ and the previous estimate is less than a given threshold, the algorithm terminates. Otherwise, another iteration commences by going back to Step 2.

The focus of this method is in iteratively adapting the weights based on the LP residual. For the first iteration, when there is no residual, unity weights are used. However, experiments show that using other types of window weights (Hamming or triangular) for the first iteration does not change the final solution. Steps 3 to 6 involve the covariance matrix of the residual $e(n)$ and the update of the weights. If all the elements of the covariance matrix are calculated, a matrix of weights equal to the inverse of the covariance matrix is obtained. The resulting weighted distance measure corresponds to a Mahalanobis distance [14]. The use of this measure effectively whitens or decorrelates the LP residual and corresponds to a Euclidean distance in the decorrelated error domain. Therefore, the noisy components in the speech signal continue to get decorrelated and are better reflected in the LP residual so that the vocal tract information is better represented in $A(z)$. In practice, we adhere to a diagonal approximation of the covariance matrix to facilitate computation of the inverse and to avoid convergence problems that were experienced by using a complete covariance matrix. The latter problem can be attributed to the fact that we are forced to estimate $N(N+1)/2$ covariance terms from a frame of only N error samples. The diagonal entries of the covariance matrix are estimated in an unbiased fashion in that there are no zeros appended to the error sequence $e(n)$ [14]. In many frames of speech, the eigenvalue spread was found to be rather high, and this led to an oscillation of the predictor coefficients. By constraining the maximum eigenvalue spread to be 100, these oscillations were removed, and faster convergence was obtained. A final implementational modification concerned the smoothing of the covariance matrix. If $\boldsymbol{C}_0$ is the covariance matrix calculated for the present iteration and $\boldsymbol{C}_{-1}$ is the covariance matrix of the previous iteration, the smoothed matrix is $\boldsymbol{C} = 0.5\boldsymbol{C}_0 + 0.5\boldsymbol{C}_{-1}$. It is this $\boldsymbol{C}$ that is used to get the weights and to represent $\boldsymbol{C}_{-1}$ for the next iteration. This step was taken to enhance the performance when speech is corrupted by Gaussian noise (elaborated on later). Since this overall procedure is iterative, it is computationally more expensive than the conventional WMSEC approach. In fact, the complexity is approximately the number of iterations times the complexity of the WMSEC approach. We have observed that the number of iterations for this approach is much lower (usually 10 to 20 times lower) than that needed for the WLAV method.

The weights are obtained to emphasize the smaller LP residual samples more and downplay the larger LP residual samples corresponding to the pitch pulses. This concept of assigning the weights to be a function of the LP residual was introduced in [15]. In our implementation, we compute the weights to achieve a Mahalanobis distance. A noniterative method to determine the weights based on the short-time energy of the speech is the subject of [7]. Since no LP residual is present, the use of the short-time energy of the speech attempts to anticipate which error samples to downplay and which to emphasize. The tradeoff is between 1) allowing for iterations and working with the true LP residual and 2) obtaining the solution in one shot and anticipating the nature of the residual. Note that the short-time energy is calculated over a window of duration thta is less than half a pitch period. For noisy speech, estimating the pitch period to set the window length is difficult.

Another iterative approach based on maximum *a posteriori* (MAP) estimation has been proposed in [16]. This is partly motivated by the observation that in the absence of noise, minimizing an L_2 norm is equivalent to a MAP estimation of $\boldsymbol{a}$, assuming that the excitation to $1/A(z)$ is white Gaussian noise [16]. When noise is present, the MAP estimate of $\boldsymbol{a}$ is found to be a nonlinear problem. This is circumvented by an approach that, at each iteration, estimates the clean speech and the predictor coefficients by solving systems of linear equations. Our iterative approach presented above differs from this method in that the following apply:

1) No assumption is made about the statistics of the excitation to $1/A(z)$.
2) No attempt is made to estimate the clean speech.
3) Successive estimates of $\boldsymbol{a}$ are based exclusively on the second-order statistics of the true LP residual.

D. Weighted Total Least-Squares Solution

In describing this approach, (1) is rewritten in vector form by considering the speech samples from a time index of n_1 to n_2 as

$$g = \boldsymbol{H}\boldsymbol{a} + \boldsymbol{e} \tag{7}$$

where $\boldsymbol{g} = [s(n_1)\, s(n_1+1) \cdots s(n_2)]^T, \boldsymbol{e} = [e(n_1)\, e(n_1+1) \cdots e(n_2)]^T$, and

$$\boldsymbol{H} = \begin{bmatrix} s(n_1-1) & s(n_1-2) & \cdots & s(n_1-p) \\ s(n_1) & s(n_1-1) & \cdots & s(n_1-p+1) \\ \vdots & \vdots & \cdots & \vdots \\ s(n_2-1) & s(n_2-2) & \cdots & s(n_2-p) \end{bmatrix}. \tag{8}$$

The methods considered so far assume that the error only occurs in the observed speech samples denoted by $\boldsymbol{g}$ and minimizes various objective functions based on the error.

We deviate from this approach by assuming that errors can occur both in $\boldsymbol{g}$ and $\boldsymbol{H}$ to get

$$\boldsymbol{g} - \boldsymbol{e} = (\boldsymbol{H} - \boldsymbol{E})\boldsymbol{a} \tag{9}$$

or equivalently

$$\{[\boldsymbol{g}|\boldsymbol{H}] - [\boldsymbol{e}|\boldsymbol{E}]\}[-1|\boldsymbol{a}]^T = 0 \tag{10}$$

where $|$ denotes augmentation. The statement of the total least-squares problem is to minimize $\|[\boldsymbol{e}|\boldsymbol{E}]\|_F$, where F

denotes the Frobenius norm subject to the constraint of (10) [17]. Weighting can be introduced by premultiplying and postmultiplying $[\boldsymbol{e}|\boldsymbol{E}]$ by nonsingular, diagonal matrices $\boldsymbol{D}$ and $\boldsymbol{T}$, respectively [18]. Then, the Frobenius norm of $\boldsymbol{D}[\boldsymbol{e}|\boldsymbol{E}]\boldsymbol{T}$ is minimized subject to (10). The total least-squares approach has been used in the context of estimating closely spaced frequencies of multiple sinusoids in the presence of noise [19]. It is our aim to see if this formulation leads to robust estimation of the predictor coefficients when speech is corrupted by noise.

For obtaining $\boldsymbol{a}$, the first step is to do a singular value decomposition of $\boldsymbol{D}[\boldsymbol{e}|\boldsymbol{E}]\boldsymbol{T}$ as $\boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices, and $\boldsymbol{B}$ is a diagonal matrix containing the singular values in descending order. Consider the case when the smallest singular value does not repeat. Let the singular vector in $\boldsymbol{V}$ that corresponds to the smallest singular value be $\boldsymbol{v}_{p+1} = [v_{p+1}(1)\, v_{p+1}(2) \cdots v_{p+1}(p+1)]^T$. Then

$$\boldsymbol{a} = -\frac{1}{v_{p+1}(1)t(1)}[t(2)v_{p+1}(2)\ t(3)v_{p+1}(3) \cdots t(p+1)v_{p+1}(p+1)]^T \tag{11}$$

if $v_{p+1}(1) \neq 0$, where the $t(i)$ are the diagonal elements of $\boldsymbol{T}$. If the smallest singular value repeats, all the corresponding singular vectors in $\boldsymbol{V}$ can be considered to evaluate a set of candidate vectors $\boldsymbol{a}$. The selected vector $\boldsymbol{a}$ corresponds to that which gives the smallest norm (either L_2 or L_1) of the LP residual. Note that this has never occurred in practice. In our work, we set the diagonal elements of $\boldsymbol{D}$ to be the Hamming window weights. Various ways to choose $\boldsymbol{T}$ were attempted for an even order p. If the diagonal elements of $\boldsymbol{T}$ are symmetric about a center coefficient, all of the roots of $A(z)$ occur on the unit circle. To alleviate this, the symmetry in $\boldsymbol{T}$ was destroyed by setting its diagonal elements to be the Hamming window weights multiplied by an asymmetric response that has a maximum at the peak of the Hamming window. The indices are $n_1 = 1$ and $n_2 = N$. The acronym for this method is WTLS.

III. Experimental Results

In this section, we discuss the results obtained by the following experiments that serve to compare the various LP methods described above. First, the time domain nature of the LP residual is examined. Second, we enumerate the number of instances that lead to a nonminimum-phase LP polynomial $A(z)$ and describe the implications of such a phenomenon. Third, two different types of noise are added to the speech, namely, white Gaussian noise and impulsive noise. A performance analysis of how the LP polynomial deviates from what is obtained for clean (noiseless) speech is given. In the next section, results of speaker identification experiments using the different LP analysis methods are given. The TIMIT database consisting of speakers from the New England area was used in the experiments. The speech is low-pass filtered to 3.4 kHz and downsampled to a sampling rate of 8 kHz. For each LP method, the order of prediction is $p = 12$, and each analysis frame is 30 ms in duration. A 20-ms overlap between successive analysis frames is used. When preemphasis is applied, the speech is passed through a nonrecursive filter $1 - 0.95z^{-1}$.

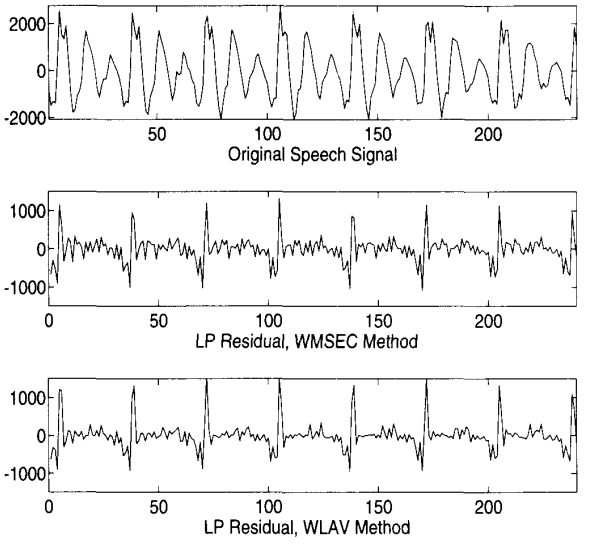


Fig. 1. Waveform plots of original speech and LP residuals formed by the WMSEC and WLAV methods.

A. LP Residual

Here, we compare the LP residual obtained by the conventional WMSEC method with that obtained for the WLAV, IWLS, and the WTLS approaches. In Fig. 1, waveform plots of a frame of speech with the residuals obtained by the WMSEC and WLAV methods are shown. The pitch pulses in the residual formed by the WLAV approach are generally of a higher amplitude than the pitch pulses in the residual resulting from the WMSEC method. This is because the pitch pulses are outliers that are more ignored by the WLAV method than the WMSEC method. Fig. 2 depicts waveform plots of the same frame of speech with the residuals obtained by the WMSEC and WTLS methods. There is little difference between the two residuals. From a computational point of view, solving a system of equations by the Cholesky decomposition (for the WMSEC approach) is more efficient than doing a singular value decomposition (for the WTLS method). Fig. 3 shows waveform plots of the speech and the residuals obtained by the WMSEC and IWLS methods. The IWLS method downplays the large-amplitude pitch pulses for the prediction and, hence, leads to a residual with higher amplitude pitch pulses than its WMSEC counterpart. The residuals for the WLAV and the IWLS approaches are similar. However, for this frame, only 10 iterations are needed for the IWLS method as compared with 245 iterations for the WLAV method. This glaring disparity in the number of iterations for the two techniques is consistent for many frames of speech.

B. Minimum Phase Property

The minimum phase property of $A(z)$ is only guaranteed by using the WMSEA approach. For applications in which the signal is reconstructed (like in speech coding), this property

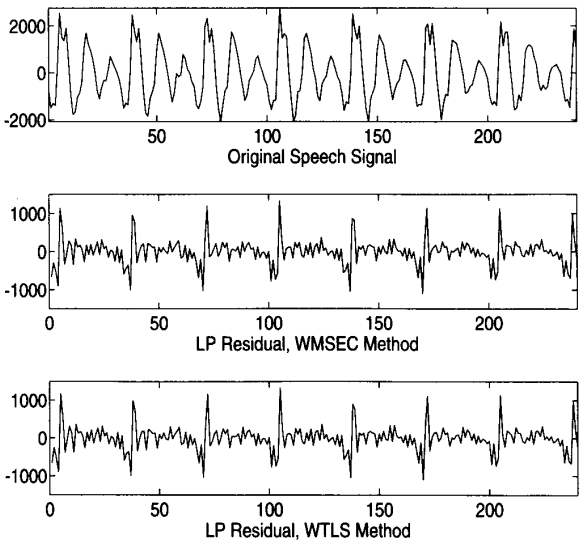


Fig. 2. Waveform plots of original speech and LP residuals formed by the WMSEC and WTLS methods.

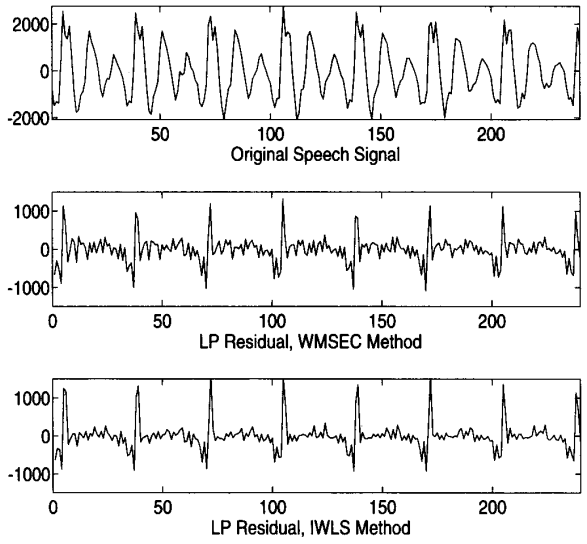


Fig. 3. Waveform plots of original speech and LP residuals formed by the WMSEC and IWLS methods.

is important. In the case of speaker identification in which no signal reconstruction takes place, this can seem to be rather insignificant. However, the LP cepstrum, which is the most common feature used for speaker identification, is affected as follows. If $A(z)$ is minimum phase, the LP cepstrum is a causal sequence. For a mixed-phase $A(z)$, the LP cepstrum is a two-sided sequence. The zeros of $A(z)$ inside the unit circle contribute to the causal component of the cepstrum, and those outside the unit circle contribute to the anticausal component [20]. Since a mixed phase $A(z)$ does not occur for every frame of speech, there is some ambiguity as to which cepstral coefficients should be considered for speaker identification. Initially, the idea of considering certain cepstral coefficients for the mixed-phase case and certain coefficients for the minimum-phase case was conceived. However, experiments revealed that for all the LP methods, $A(z)$ is minimum phase about 98% of the time. Since the mixed phase case is rather rare, we continue our experiments by taking the zeros of $A(z)$ outside the unit circle and reflecting them inside.

C. Noise Addition

We add noise to the speech and implement the methods for both the clean (noiseless) and the noisy speech. The objective is to get a set of parameters for noisy speech that are close to that of the clean speech. To compare the methods in terms of this aim, we consider the following performance measures. Let $\boldsymbol{a}_c$ be the vector of predictor coefficients for clean speech and $\boldsymbol{a}_n$ be the vector of predictor coefficients for noisy speech. The predictor coefficient signal-to-noise ratio (SNR), in decibels, is defined to be

$$\text{Predictor Coefficient SNR} = 20 \log \frac{\|\boldsymbol{a}_c\|_2}{\|\boldsymbol{a}_c - \boldsymbol{a}_n\|_2}. \quad (12)$$

The predictor coefficients are converted to a set of cepstral coefficients, and the cepstrum SNR is similarly defined. The first 12 cepstral coefficients are used for the study. As mentioned earlier, it is the cepstrum that is used for speaker identification. Moreover, the squared L_2 norm of the difference between two cepstral vectors taken from clean and noisy speech is a reflection of the mean-square distance between the log spectra of the clean and noisy speech [21]. Given $\boldsymbol{a}_c$, the poles of $1/A(z)$ are found, and the four complex poles that have the largest magnitude are selected as they represent the formant frequencies. These four poles are compared with the corresponding complex poles derived from $\boldsymbol{a}_n$, and a formant pole SNR is determined. The corresponding poles are those closest to the four formant poles of clean speech. Given a set of poles f_k of $1/A(z)$ within the unit circle, the cepstrum $c(n)$ can be shown to be given by [20]

$$c(n) = \frac{1}{n} \sum_k f_k^n. \quad (13)$$

We get a modified cepstrum from this formula by considering only the four formant poles and then calculating a modified cepstrum SNR. Note that the formant pole SNR and the modified cepstrum SNR are somewhat unrealistic in that the four poles of largest magnitude for clean speech do not necessarily correspond to the largest magnitude poles for noisy speech. Therefore, doing a pole selection on noisy speech without the knowledge of the poles for clean speech and using the modified cepstrum as a feature for speaker identification is risky. However, we evaluate these measures to get an idea of how the formant poles of clean speech migrate in the presence of noise. For each frame, the various SNR values are different, and hence, what we present are average SNR values taken over 42 556 voiced analysis frames. These frames come from 200 sentences spoken by 20 speakers (ten sentences per speaker).

In the first experiment, we add white Gaussian noise to the clean speech corresponding to SNR values of 30, 20, 10, and 5 dB. With preemphasis, the noise becomes colored and is

TABLE I
PREDICTOR COEFFICIENT SNR IN DECIBELS (GAUSSIAN NOISE). THE TWO VALUES FOR EACH METHOD ARE FOR THE CASES OF NOT USING AND USING PREEMPHASIS

SNR of noisy speech	LP Method									
	WMSEA		WMSEC		WLAV		IWLS		WTLS	
30 dB	19.89	19.31	20.40	20.01	15.61	15.09	17.23	16.85	19.48	19.10
20 dB	11.84	11.23	11.96	11.53	9.76	9.08	10.13	9.53	11.56	11.10
10 dB	7.10	6.10	7.04	6.22	6.23	5.09	6.37	5.20	6.85	6.04
5 dB	5.79	4.57	5.72	4.67	5.29	3.80	5.43	3.88	5.58	4.54

TABLE II
CEPSTRUM SNR IN DECIBELS (GAUSSIAN NOISE). THE TWO VALUES FOR EACH METHOD ARE FOR THE CASES OF NOT USING AND USING PREEMPHASIS

SNR of noisy speech	LP Method									
	WMSEA		WMSEC		WLAV		IWLS		WTLS	
30 dB	23.06	20.87	23.57	21.60	17.93	16.00	19.84	18.07	22.65	20.69
20 dB	13.52	11.58	13.74	11.95	11.08	9.11	11.55	9.62	13.34	11.50
10 dB	7.11	5.25	7.21	5.42	6.07	4.08	6.22	4.20	7.08	5.25
5 dB	5.16	3.29	5.24	3.43	4.47	2.44	4.61	2.54	5.18	3.32

TABLE III
FORMAT POLE SNR IN DECIBELS (GAUSSIAN NOISE). THE TWO VALUES FOR EACH METHOD ARE FOR THE CASES OF NOT USING AND USING PREEMPHASIS

SNR of noisy speech	LP Method									
	WMSEA		WMSEC		WLAV		IWLS		WTLS	
30 dB	36.30	36.72	37.21	37.67	29.82	31.18	32.84	34.41	36.39	36.80
20 dB	27.38	28.11	27.89	28.67	24.08	25.58	25.27	26.90	27.44	28.13
10 dB	21.03	21.90	21.32	22.22	19.55	21.01	20.03	21.55	21.08	21.91
5 dB	18.73	19.79	19.05	20.07	17.68	19.24	18.12	19.72	18.86	19.81

TABLE IV
MODIFIED CEPSTRUM SNR IN DECIBELS (GAUSSIAN NOISE). THE TWO VALUES FOR EACH METHOD ARE FOR THE CASES OF NOT USING AND USING PREEMPHASIS

SNR of noisy speech	LP Method									
	WMSEA		WMSEC		WLAV		IWLS		WTLS	
30 dB	25.25	24.30	25.79	25.08	18.50	18.91	21.48	22.06	25.09	24.23
20 dB	16.45	15.75	16.61	16.14	12.75	13.30	13.88	14.53	16.26	15.63
10 dB	10.27	9.61	10.22	9.76	8.22	8.70	8.64	9.15	10.08	9.48
5 dB	8.11	7.53	8.06	7.63	6.43	6.96	6.77	7.31	7.98	7.42

more dominant at the high frequencies. Tables I to IV show the performance values for the LP analysis methods. The WMSEA and WMSEC methods show the best performance. This is not surprising since an L_2 estimate is the maximum likelihood estimate for a Gaussian error density. The WTLS method is only slightly inferior to the L_2 approaches. The WLAV method shows the lowest SNR values. The IWLS method would show an equivalent performance if termination occurred at the very first iteration. However, the iterative character was introduced to combat different types of noise (the condition of impulse noise is discussed later). If no smoothing of the covariance matrix was done, the number of iterations increased, and a performance worse than the WLAV method resulted. By introducing smoothing, the convergence was accelerated, and the performance was enhanced. Finally, note that the disparity in the performance is less as the SNR of the noisy speech is reduced. By introducing more noise at the high frequencies where the speech spectrum is generally more attenuated, preemphasis diminishes the predictor coefficient SNR and cepstrum SNR. However, the formant pole SNR and modified cepstrum SNR can increase by using preemphasis. Fig. 4 shows histograms of the cepstrum SNR (with preemphasis) for speech corrupted by white Gaussian noise with an SNR of 20 dB.

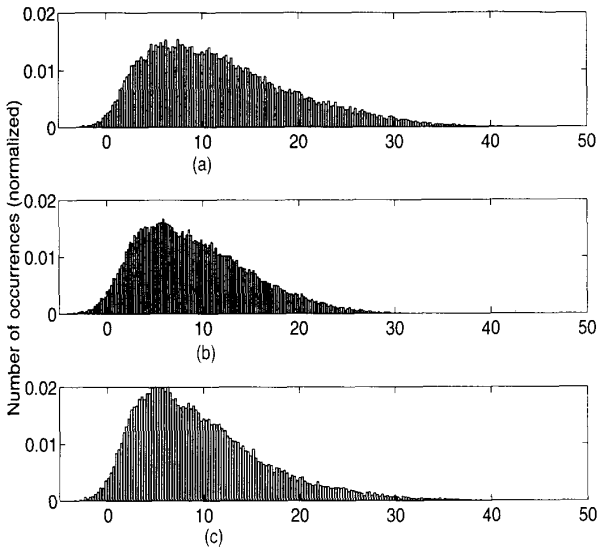


Fig. 4. Histogram of cepstrum SNR values (with preemphasis) for three methods given speech corrupted by white Gaussian noise with an SNR of 20 dB: (a) WMSEC method; (b) WLAV method; (c) IWLS method.

The second experiment involves the addition of impulsive noise. The speech signal is divided into blocks of 10 ms. For each block, an impulse of noise is injected as follows. The maximum of the absolute value of the signal amplitude is calculated. An impulse of this maximum amplitude is added to the speech at a random sample location. The sign of this impulse is the same as that of the speech sample to which it is added. For each frame of 30 ms, there are three impulses, thereby bringing in a heavy noise content. Table V shows the performance results for this case of impulsive noise. The IWLS approach shows the highest values of the predictor coefficient SNR and the cepstrum SNR. The WLAV method is the second best in this regard. Impulse noise samples present themselves as outliers in the speech signal and are better ignored by the IWLS and WLAV methods. The inferior performance of the WMSEA and WMSEC methods suggests that stopping the IWLS algorithm at the first iteration is bad for speech corrupted by impulse noise. In fact, when no smoothing of the covariance matrix took place, the performance of the IWLS approach improved. Therefore, the smoothing compromised the performance for impulse noise but enhanced the performance for Gaussian noise. However, the diminished performance for impulse noise does not sacrifice the claim of IWLS being the most robust and simultaneously imposes less of a computational burden. Regarding the formant poles and the modified cepstrum, the WLAV and the IWLS methods are usually the best. However, the differences in the formant pole SNR are not much. This implies that impulse noise causes a similar movement of the formant poles for all of the methods. However, as mentioned earlier, locating the formant poles of noisy speech is a very difficult problem. Preemphasis causes a significant decrease in the cepstrum SNR and a slight decrease in the modified cepstrum SNR. Fig. 5 shows histograms of the cepstrum SNR (no preemphasis) for speech corrupted by impulse noise.

TABLE V
VARIOUS SNR VALUES IN DECIBELS (IMPULSIVE NOISE). THE TWO VALUES FOR EACH METHOD ARE FOR THE CASES OF NOT USING AND USING PREEMPHASIS

Performance measure	LP Method									
	WMSEA		WMSEC		WLAV		IWLS		WTLS	
Predictor coefficient SNR	1.26	1.24	1.19	1.24	2.21	2.02	2.48	2.21	1.13	1.21
Cepstrum SNR	4.51	2.10	4.39	2.09	6.75	3.99	7.03	4.44	4.31	1.98
Formant pole SNR	17.14	17.08	17.03	17.05	18.01	17.99	17.67	17.59	16.76	17.00
Modified cepstrum SNR	8.37	6.71	8.00	6.57	8.56	7.81	8.21	7.52	7.90	6.52

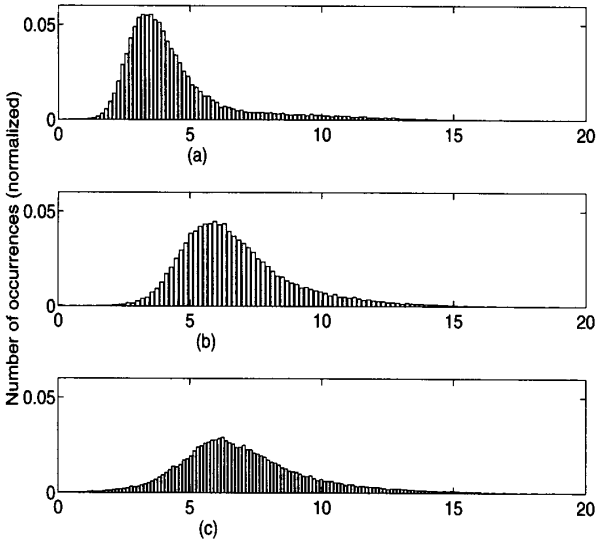


Fig. 5. Histogram of cepstrum SNR values (no preemphasis) for three methods given speech corrupted by impulse noise: (a) WMSEC method; (b) WLAV method; (c) IWLS method.

IV. SPEAKER IDENTIFICATION

A closed-set speaker identification experiment was performed using 20 speakers. The objective is to compare the identification success rates of each of the LP methods. A vector quantizer (VQ) classifier [22] is used in the following manner. For each speaker, the voiced portions of the first five sentences in the TIMIT database are used to extract a set of 12-dimensional cepstral vectors. This comprises a training set from which a VQ codebook is designed. The learning VQ design procedure (LVQ1) as described in [23] is used. The LVQ1 algorithm is formulated to take a training set consisting of cepstral vectors from all speakers and designing a codebook that has a label pointing to a particular speaker for each entry. This will not guarantee that each speaker is represented by the same number of codebook entries. Therefore, as done in [22] and [24], a total of 20 codebooks are designed, each corresponding to an individual speaker. For the design of a particular codebook, only the training data for the corresponding speaker is used, thus providing one default label. Note that if a nonminimum phase $A(z)$ was not modified to be minimum phase, two VQ codebooks for each speaker would be required. One codebook would be for the minimum phase case in which the first 12 components of the causal cepstrum are represented. The other codebook is for the mixed phase case in which six anticausal components and six causal components of the cepstrum are used. However, as mentioned

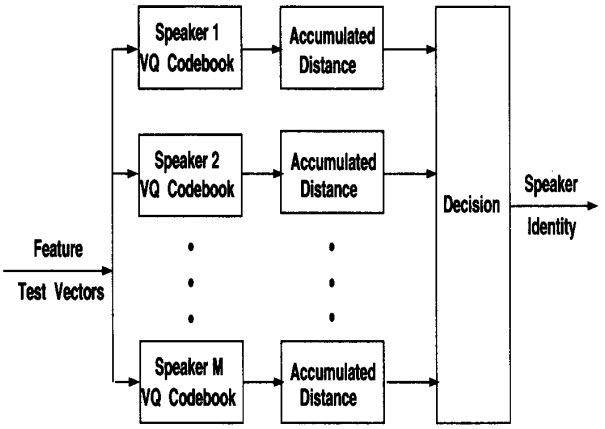


Fig. 6. Block diagram of speaker identification system.

earlier, a mixed-phase $A(z)$ is rare, thereby providing little training data and hence contributing very little to the overall system. Therefore, the idea of forcing a minimum-phase $A(z)$ is beneficial for the speaker identification problem. Note that only clean speech is used for training. In addition, the speech is preemphasized.

The remaining five sentences for each speaker are used individually for testing, thereby giving a total of 100 test sentences. Again, the speech is preemphasized. For each test sentence, a cepstral vector in a voiced frame is quantized by each of the 20 codebooks. The quantized vector is that which is closest in Euclidean distance to the test vector. For every voiced frame, 20 different distances are recorded: one for each codebook. These distances are accumulated over the entire test sentence. The codebook that renders the smallest accumulated distance identifies the speaker. The identification success rate is enumerated over the 100 test sentences. Fig. 6 shows a block diagram of the system.

Each LP method gives different training data, test data, and VQ codebooks. The aim is to compare the various LP methods. To do this, three experiments are carried out. First, clean speech is used for testing. Second, speech corrupted by additive Gaussian noise is used for testing. Third, the test speech is degraded by impulse noise.

Consider the case when the test speech is clean. Table VI shows the success rates as a function of the codebook size. As the codebook size increases, the performance of each method tends to saturate. Therefore, merely using a large codebook size does not benefit in terms of performance and imposes a cost in terms of memory and search complexity. In the limit, as the codebook size equals the number of vectors in the training set, a nearest-neighbor classifier is obtained. Experiments have shown that the nearest-neighbor classifier is inferior to the VQ technique using modest size codebooks [24]. This is because overlearning of the training data has taken place. The saturation point for all the methods is for a codebook size of 64 for which the performance is about the same for all the LP techniques. For a codebook size of 32 (which is practically more feasible), the proposed WLAV, IWLS, and WTLS methods are slightly superior to the conventional WM-

TABLE VI
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR CLEAN SPEECH

Codebook size	LP Method				
	WMSEA	WMSEC	WLAV	IWLS	WTLS
16	91	91	90	89	89
32	92	93	97	95	95
64	95	95	95	95	96
128	96	96	96	96	96
256	96	96	96	95	97

TABLE VII
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR SPEECH DEGRADED BY WHITE GAUSSIAN NOISE. THE CODEBOOK SIZE IS 32.

SNR of noisy speech	LP Method				
	WMSEA	WMSEC	WLAV	IWLS	WTLS
30	83.5	89	90	86	83.5
20	57.5	54	51	44.5	51.5
10	17.5	15	22	16.5	21
5	14	7	10	8.5	16

SEA and WMSEC approaches. However, the 95% confidence intervals [25] show much overlap, thereby indicating that the performance of the LP methods is practically the same.

When the test speech is corrupted by white Gaussian noise, the identification success rate and the cepstrum SNR are somewhat incompatible. To check for consistency, we repeated the experiment using Gaussian noise emanating from a different initial random number generator seed. In Table VII, we present the results using 200 test utterances (since the experiment is done twice) and a codebook of size 32. When the SNR of the noisy speech is 30 dB, the WLAV method shows the best success rate, whereas the WMSEA method shows the least success rate. This is in direct contrast to the cepstrum SNR in which the WMSEA method is better. However, the 95% confidence intervals have much overlap. The lower limit for the WLAV method is about 84% [25]. The upper limit for the WMSEA approach is about 90% [25]. The same type of observations can be made when comparing the WLAV and IWLS techniques when the SNR of the noisy speech is 20 dB. The situation again repeats when comparing the WMSEC and WTLS approaches when the SNR of the noisy speech is 5 dB. For the case of Gaussian noise, the success rates of the LP methods are comparable. The relative success rates cannot be predicted based on the cepstrum SNR.

Consider the case when the test speech is degraded by impulse noise. In Table VIII, we present the results using 100 test utterances and a codebook of size 32. For impulse noise, the success rates are indeed compatible with the cepstrum SNR when comparing the LP approaches. The 95% confidence interval for the WMSEA, WMSEC, and WTLS approaches ranges from 5 to 11 [25]. The 95% confidence interval for the WLAV method is from 10 to 27. This clearly demonstrates the superiority of the WLAV method. The 95% confidence interval for the IWLS technique is from 17 to 35. Although there is overlap with the interval of the WLAV method, the IWLS method is still preferred due to computational considerations. This experiment was repeated using a distortion measure equal to 1 minus the cosine of the angle between two cepstral vectors [26]. This distortion measure was proposed as being more robust than the Euclidean distance [26]. The success rate for the WMSEA, WMSEC, and WTLS methods increased to 11%.

TABLE VIII
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR SPEECH DEGRADED BY IMPULSE NOISE. THE CODEBOOK SIZE IS 32

LP Method				
WMSEA	WMSEC	WLAV	IWLS	WTLS
6	6	17	24	7

The WLAV and IWLS approaches showed a performance of 28 and 29%, respectively. Although robust LP analysis is important in alleviating the mismatch between the training and testing conditions, it is not the sole robust component of a speaker identification system.

Speaker identification is usually done by preemphasizing the speech. In the absence of preemphasis, the performance of all the methods (when clean speech is used as the test material) is essentially the same with the exception of the WTLS approach. For the WTLS procedure, the performance deteriorates by about 10%.

V. SUMMARY AND CONCLUSIONS

In this paper, three linear predictive analysis methods are formulated to achieve a clear separation of the vocal tract information and the pitch information in the speech signal. In addition, robustness to noise is desired. A comparison with the conventional least-squares solution manifested through an autocorrelation (WMSEA) and covariance (WMSEC) formulation is made. The WLAV method is based on minimizing the weighted least absolute value, puts relatively less emphasis on the outliers, and offers a solution based on the iterative simplex method of linear programming. The IWLS approach iteratively updates the weights for a least-squares analysis in order to minimize a Mahalanobis distance. Both the WLAV and the IWLS approaches deemphasize the larger residual samples. Consequently, the pitch pulses in the LP residual are more apparent when using these two methods. The WTLS approach assumes an error in both the observations and the data and minimizes a Frobenius norm. A one-shot solution is obtained by a singular value decomposition, which is slightly more complex than the conventional approaches that merely involve the solution of a system of equations.

White Gaussian noise and impulse noise are added to the speech in order to investigate the robustness of the methods. We study how the LP parameters computed for noisy speech deviate from that found for clean speech. The predictor coefficients and the cepstrum (which is used for speaker identification) are the parameters considered. For Gaussian noise, the WMSEA and the WMSEC approaches are the best. However, the disparity in the performance diminishes as the SNR decreases. In the case of impulse noise, the IWLS and the WLAV methods are the best.

In the context of speaker identification, a VQ classifier for 20 speakers was trained using clean speech. It was found that a small but effective codebook size is 32. For this size, the performance of the various LP approaches are comparable when the test speech material is either clean or corrupted by white Gaussian noise. When the test speech material is degraded by impulse noise, the relative performance of the

LP methods is directly related to the relative deviations in the cepstral vectors.

A clean speech signal corresponding to a sustained vowel (/a/) spoken by a male was processed by each of the LP methods. The locations of the four formants are known, and the objective is to compare the formant poles obtained by the LP methods. It is revealed that the mean value of the four formant poles (taken over 98 frames) are practically the same for each of the LP methods. Moreover, the standard deviations are low and about the same for the LP methods. The maximum standard deviation is for the third formant pole and is equal to 0.05.

ACKNOWLEDGMENT

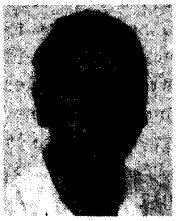
The LVQ software design package prepared by the LVQ programming team of the Helsinki University of Technology was used to design the codebooks for the speaker identification system. The authors thank Dr. S. Yu of nCUBE corporation for generating the software for running the WLAV and IWLS methods on the nCUBE machine.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[2] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.*, vol. COM-30, pp. 600–614, Apr. 1982.
[3] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback," *IEEE J. Sel. Areas Commun.*, vol. 6, pp. 364–382, Feb. 1988.
[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, June 1974.
[5] ——, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.
[6] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 488–494, Dec. 1976.
[7] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Commun.*, vol. 12, no. 2, pp. 69–81, June 1993.
[8] R. J. Mammone, *Computational Methods of Signal Recovery and Recognition*. New York: Wiley, 1992.
[9] B. Noble and J. W. Daniel, *Applied Linear Algebra*. Englewood Cliffs, NJ: Prentice-Hall, 1977.
[10] R. J. Mammone and G. Eichmann, "Restoration of discrete Fourier spectra using linear programming," *J. Opt. Soc. Amer.*, vol. 72, pp. 987–992, Aug. 1982.
[11] R. J. Rothacker, R. J. Mammone, and S. Davidovici, "Spectrum enhancement using linear programming," *IEEE Int. Conf. Acoust., Speech Signal Processing* (Tokyo, Japan), Apr. 1986, pp. 43.10.1–43.10.4.
[12] R. J. Mammone and G. T. Sentman, "Multi-pulse LPC using linear programming in the frequency domain," *IEEE Int. Conf. Acoust., Speech ASignal Processing* (Tokyo, Japan), Apr. 1986, pp. 56.6.1–56.6.4.
[13] J. Schroeder and R. Yarlagadda, "Linear predictive spectral estimation via the L_1 norm," *Signal Processing*, vol. 17, pp. 19–29, June 1989.
[14] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[15] C.-H. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 642–650, May 1988.
[16] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing.*, vol. ASSP-26, pp. 197–210, June 1978.
[17] S. Van Huffel and J. Vandewalle, *The Total Least-Squares Problem, Computational Aspects and Analysis*. Soc. Indust. Applied Math., 1991.
[18] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *Siam J. Numer. Anal.*, vol. 17, pp. 883–893, Dec. 1980.
[19] M. A. Rahman and K.-B. Yu, "Total least squares approach for frequency estimation using linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1440–1454, Oct. 1987.
[20] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[21] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[22] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Comp. Speech Language*, vol. 22, pp. 143–157, 1987.
[23] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, pp. 1464–1480, Sept. 1990.
[24] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks versus conventional classifiers," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 194–205, Jan. 1994.
[25] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[26] D. Mansour and B.-H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 37, pp. 1659–1671, Nov. 1989.

Ravi P. Ramachandran was born in Bangalore, India, on July 12, 1963. He received the B.Eng. degree (with great distinction) from Concordia University, Montreal, Canada, in 1984 and the M.Eng. and Ph.D degrees from McGill University, Montreal, Canada, in 1986 and 1990, respectively.

From January to June 1988, he was a visiting postgraduate researcher at the University of California at Santa Barbara. From October 1990 to December 1992, he worked in the Speech Research Department at AT&T Bell Laboratories, Murray Hill, NJ. Since January 1993, he has been a Research Assistant Professor at the CAIP Center, Department of Electrical Engineering, Rutgers University, Piscataway, NJ. His main research interests are in speech processing, data communications, and digital signal processing.



Mihailo S. Zilovic was born in Belgrade, Yugoslavia, on July 26, 1961. He received the Diploma of Engineering (Electrical) degree from Belgrade University, Belgrade, Yugoslavia, the M.E.E. degree from the City College of New York, and the Ph.D. degree from the City University of New York in 1986, 1989, and 1993, respectively.

From February to June 1993, he was a part-time postgraduate researcher at the CAIP Center, Rutgers University, Piscataway, NJ, a part-time lecturer with the Electrical and Computer Engineering Department, Rutgers University, and an Adjunct Assistant Professor with the Electrical Engineering Department, the City College of New York. Since July 1993, he has been a Research Assistant Professor at the CAIP Center, Rutgers University. His main research interests are in speech processing, digital signal processing, and multidimensional system theory.

Richard J. Mammone (S'75–M'81–SM'86) received the B.E.E., M.E.E., and Ph.D. degrees from the City University of New York in 1975, 1977, and 1981, respectively.

He is currently a Professor of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ. His research and teaching interests are in the areas of image and speech processing and neural networks. He has numerous publications and patents in these areas and is a frequent consultant to industry and government agencies.

Dr. Mammone is an Associate Editor of the journal *Pattern Recognition* and was an Associate Editor of *IEEE Communications Magazine*. He is co-editor of a book on neural networks with Y. Y. Zeevi. He is a member of OSA, SPIE, Eta Kappa Nu, and Sigma Xi.