RECOGNITION OF NOISY SPEECH USING CUMULANT-BASED LINEAR PREDICTION ANALYSIS

K.K. Paliwal and M.M. Sondhi
Acoustics Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

ABSTRACT — At present, most speech recognizers use linear prediction (LP) parameters which are estimated from the speech signal through the conventional autocorrelation method of LP analysis. The performance of these recognizers deteriorates drastically for noisy speech, specially when it is not feasible to train and test these recognizers under identical noise conditions. To alleviate this problem, the LP parameters are derived here through a cumulant-based LP analysis method. It is shown that the resulting recognizer improves performance in the presence of Gaussian noise which may be white or colored.

1. INTRODUCTION

At present, most speech recognition systems use linear prediction (LP) parameters which are derived from the speech signal by matching the autocorrelation (or, the power spectrum) [1]. These systems perform well for clean speech, but their performance degrades drastically for noisy speech; specially under mismatched noise conditions (i.e., where the training and the testing are done under different noise conditions). A number of studies are reported in the literature to deal with this problem [2, 3, 4, 5, 6, 7].

In this paper, we propose to use LP parameters derived through a cumulant-based LP analysis method for the recognition of noisy speech. This method assumes that the speech signal is *non-Gaussian* and satisfies the all-pole (or, autoregressive (AR)) model. The LP parameters are estimated here by matching the all-pole model cumulants with the cumulants of the speech signal [8]. The cumulant-based LP analysis method can suppress the effects of additive *Gaussian* noise whether white or colored. This is because Gaussian processes have identically zero cumulants of all orders greater than two [8]. To avoid the problem of the large dynamic range associated with higher-order cumulants, we consider only third-order cumulants in this paper. However, the results reported in this paper can be extended to higher-order cumulants. When the recognition system uses the LP parameters derived through the cumulant-based method, it is found to be robust to white as well as colored Gaussian noise. Note that none of the studies [2, 3, 4, 5, 6, 7], mentioned above, deals with recognition in the presence of colored noise.

The organization of the paper is as follows. In Section 2, the cumulant-based LP analysis method is described and the effect of additive Gaussian noise on AR spectral estimation is demonstrated. Section 3 describes the speech recognition experiments for the case of additive white Gaussian noise; while the experiments with colored Gaussian noise are described in Section 4. Conclusions are reported in Section 5.

2. CUMULANT-BASED LP ANALYSIS METHOD

Let $\{x_n\}$ be a pth order AR process; i.e.,

$$\sum_{k=0}^{p} a_k x_{n-k} = u_n, \qquad (1)$$

where $\{a_k\}$ are the LP coefficients (with $a_0 = 1$) and $\{u_n\}$ is a *non-Gaussian* i.i.d. innovations process with $E\{u_n\} = 0$, $E\{u_n^2\} = \alpha$ and $E\{u_n^3\} = \beta$. If the AR process is ergodic, its third-order cumulants (which, for a zero-mean process are the same as its third-order moments [8]) are given by

$$\begin{aligned} R(i,j) &= E\{x_n x_{n+i} x_{n+j}\} \\ &= \sum_{n=-\infty}^{\infty} x_n x_{n+i} x_{n+j}. \end{aligned} \qquad (2)$$

These cumulants satisfy the following recursions [9]:

$$\sum_{k=0}^{p} a_k R(k-i, k-j) = \beta\delta(i,j), \quad (i,j) \geq 0, \qquad (3)$$

where $\delta(i,j)$ is the two-dimensional unit impulse function. These recursion equations can be solved to compute β and the LP coefficients $\{a_k\}$. We are interested here only in the $\{a_k\}$, which can be estimated by solving p equations defined in terms of $2p$ cumulants on the line $i = j$ [9]. It has been observed [10] that the resulting estimates of $\{a_k\}$ have large variance in practice. The estimates can be improved by using an overdetermined set of equations. In the present paper, we compute the $\{a_k\}$ from the overdetermined set of $p(p+3)/2$ recursion equations obtained from Eq. (3) by setting $1 \leq i \leq p$ and $0 \leq j \leq i$.

In order to compute the cumulants through Eq. (2), it is necessary to know the speech signal for all time; i.e., from $-\infty$ to $+\infty$. In practice, speech is analyzed on a short-time basis; i.e., only a finite segment of N speech samples, $\{s_n, n = 1, \ldots, N\}$, is available for analysis. So, as in the autocorrelation method of LP analysis [1], the speech signal has to be windowed. We have found a Hamming window to be preferable over a rectangular window for this purpose.

Analogously to the covariance method of LP analysis [1], we can formulate a "covariance-like" method for cumulant-based LP also. In this formulation, the LP coefficients are computed using the following equations:

$$\sum_{k=0}^{p} a_k C_k(i,j) = 0, \quad 1 \leq i \leq p, 0 \leq j \leq i, \qquad (4)$$

where the cumulants are given by

$$C_k(i,j) = \sum_{n=p+1}^{N} s_{n-k} s_{n-i} s_{n-j} \qquad (5)$$

The "autocorrelation-type" and the "covariance-type" of cumulant-based LP analysis methods, both result in comparable recognition performance for clean speech; but the covariance-type formulation performs better for the recognition of noisy speech under mismatched conditions. Because of this, the results reported hereafter in this paper are obtained by using the covariance-type formulation for the cumulant-based LP analysis.

The LP spectral envelope estimates from the cumulant-based LP analysis method are shown in Fig. 1 for vowel sounds /i/ and /a/. For comparison, we also show in this figure the estimates from the conventional autocorrelation method of LP analysis [1]. The cumulant-based LP analysis method shows formant structure similar to the conventional autocorrelation method, though the overall spectra from the two methods are quite different.

In order to study the effect of additive white Gaussian noise on the LP spectral estimation performance, we use machine-generated pseudo-random Gaussian numbers as white noise and add it to the clean speech signal. Estimated LP spectra for the clean and noisy speech at signal-to-noise ratio (SNR) equal to 20 dB are shown in Fig. 2 for the conventional autocorrelation method and in Fig. 3 for the cumulant-based method. It can be seen from these figures that the additive white Gaussian noise affects the LP spectral estimates for the autocorrelation method, while the spectral estimates remain relatively unaffected for the cumulant-based method.

3. SPEECH RECOGNITION EXPERIMENTS WITH ADDITIVE WHITE GAUSSIAN NOISE

In this section, we describe some speech recognition experiments where speech signal is corrupted by the addition of *white* Gaussian noise. The aim

CH2977-7/91/0000-0429 $1.00 © 1991 IEEE

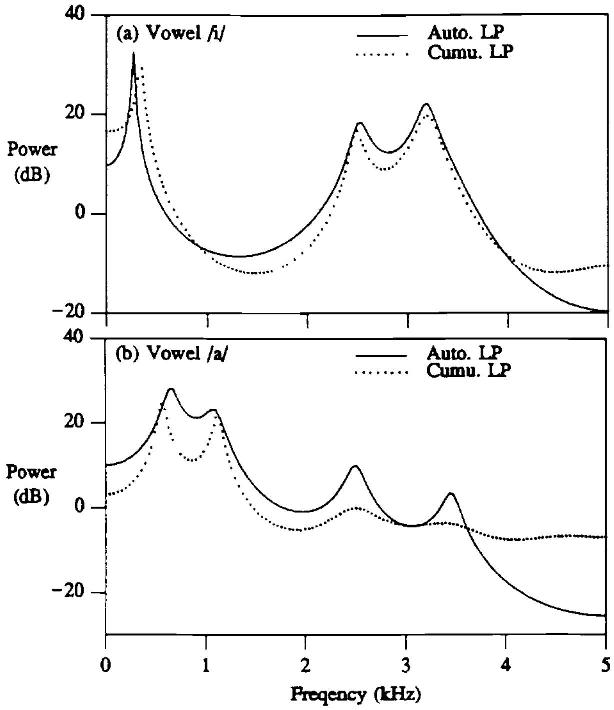

Fig. 1: LP spectral estimates of clean speech using the autocorrelation and the cumulant-based methods.

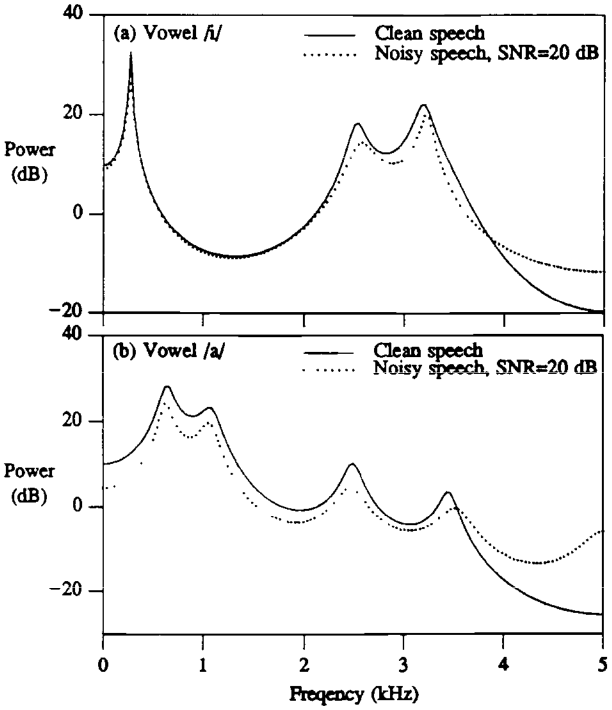

Fig. 2: LP spectral estimates of clean and noisy speech (with additive white Gaussian noise at SNR=20 dB) using the autocorrelation method.

is to study the advantage of the cumulant-based LP analysis method over the conventional autocorrelation method. As mentioned earlier, cumulant-based LP analysis assumes the speech signal to be *non-Gaussian*. We would expect this assumption to be satisfied better for vowel sounds than for non-vowel sounds (such as fricatives). For this reason, we present the performance of a vowel recognition system, in addition to that of a more general isolated word recognition system. Though the speech recognition systems were evaluated in both speaker-dependent and multi-speaker modes, we describe here only the results for the multi-speaker mode. The results for the speaker-dependent mode were found to be qualitatively similar to those for the multi-speaker mode.

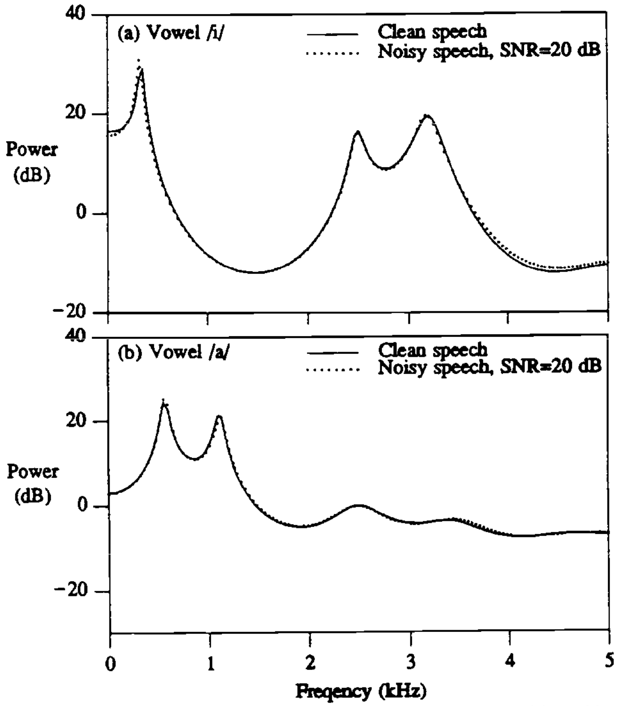

Fig. 3: LP spectral estimates of clean and noisy speech (with additive white Gaussian noise at SNR=20 dB) using the cumulant-based method.

3.1. Vowel Recognition Experiments

In these experiments, the recognition task is to classify steady-state vowel segments into 10 vowel classes. The speech data base used for this purpose is derived from 900 utterances which consist of 30 repetitions of 10 different /b/-vowel-/b/ syllables spoken by three speakers (two males and one female). These utterances were lowpass filtered to 4 kHz and digitized at 10 kHz sampling rate. The steady-state part of the vowel segment was manually located for each of the 900 utterances and a 20 ms segment excised from its center. A 10-th order LP analysis was performed for each such 20 ms segment. The first 15 repetitions from each of the three speakers were pooled together to get 45 repetitions for training. The remaining 45 repetitions from the three speakers were used for testing.

The maximum likelihood (ML) classifier is used here for vowel classification. The ML classifier classifies the input vector x (having 10 cepstral coefficients as its components) into vowel class i if $p(\mathrm{x}|i) > p(\mathrm{x}|j)$ for all $j \neq i$, where $p(\mathrm{x}|i)$ is the probability density function (or, likelihood function) for class i. In the present study, the class-conditional likelihood functions are assumed to be multivariate Gaussian (with diagonal covariance matrix).

Fig. 4 shows the vowel recognition performance for different SNRs for both the cumulant-based LP analysis method and the conventional autocorrelation method. It can be seen from this figure that the recognition performance degrades drastically with decrease in SNR for the autocorrelation method. At SNR=15 dB, the deterioration in recognition accuracy is about 32%. For the cumulant-based LP analysis method, the recognition performance remains unaffected over a wide range of SNR values. However, unfortunately, for clean speech the error rate is significantly higher than that obtained with the autocorrelation method. We do not, at present, fully understand why this is the case. One reason may be that because of poor conditioning of the matrix C in Eq. (4), the cumulant-based estimate has a large variance. In that case it might be advantageous to use *both* estimates (the cumulant-based and the autocorrelation-based) as two inde-

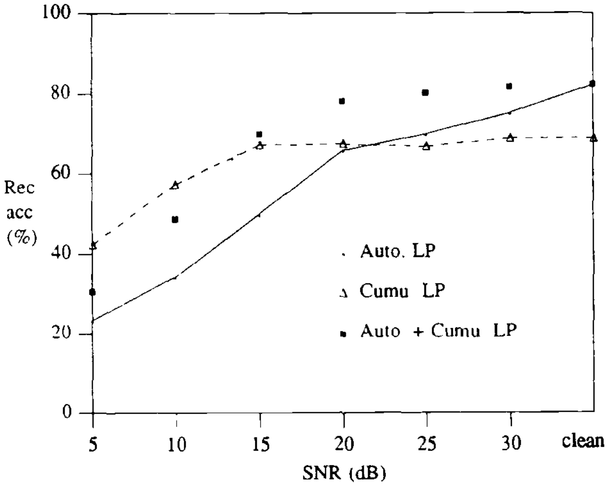

Fig. 4: Recognition accuracy for the vowel recognition task as a function of SNR for additive white Gaussian noise.

pendent measurements. When the cepstral coefficients from the two methods are combined in this manner, we get the results shown as the dotted line in Fig. 4. We see that in this case the recognition performance is comparable to that from the autocorrelation method alone for clean speech, and better for noisy speech at all SNRs.

3.2. Isolated Word Recognition Experiments

In these experiments, the recognition task is to recognize isolated words from a limited vocabulary. A hidden Markov model (HMM) based recognizer is used for this purpose [11]. The HMM is a left-to-right Bakis-type model and has five states. Single multivariate Gaussian functions (with diagonal covariance matrices) are used to characterize the probability density functions of different states. The Viterbi algorithm is used for training as well as for testing the recognizer.

In order to study the recognition performance of the cumulant-based LP analysis method and the conventional autocorrelation method, the following two different vocabularies are used: 1) the digits vocabulary consisting of 10 English digits (0-9), and 2) the alpha-digits vocabulary consisting of 39 English alpha-digits (26 alphabets (A-Z) + 10 digits (0-9) + 3 command words 'stop', 'error' and 'repeat'). The data base consists of speech from 4 talkers (2 males and 2 females). 24 utterances of each word from these 4 talkers were used for training and an additional 40 utterances for testing. The training and testing tokens were recorded over local dialed-up telephone lines, and digitized at a sampling rate of 6.67 kHz. An 8-th order LP analysis was performed every 15 ms with a frame width of 45 ms, and each frame represented in terms of 12 cepstral coefficients [12]. Endpoints of each utterance were manually determined.

The speech recognition performance was studied for noisy speech at several SNRs. The results are shown in Fig. 5 for the digits vocabulary and in Fig. 6 for the alpha-digits vocabulary. These figures include recognition results from the autocorrelation method, the cumulant-based method and the autocorrelation method + the cumulant-based method. As in the vowel recognition experiments, the recognition performance with the autocorrelation method degrades rapidly with a decrease in SNR, while the performance with the cumulant-based method remains the same over a wide range of SNR values. However, as before, the recognition performance for clean speech is quite poor. Use of cepstral coefficients from both the autocorrelation and the cumulant-based methods results in better recognition performance at all SNRs than that obtained by using the autocorrelation method alone.

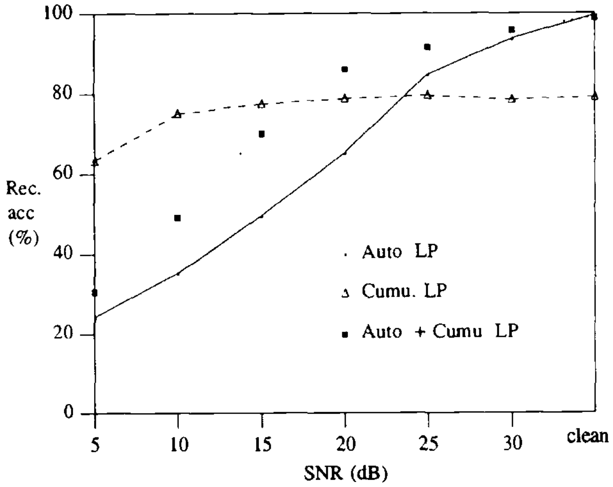

Fig. 5: Recognition accuracy for the digit recognition task as a function of SNR for additive white Gaussian noise.

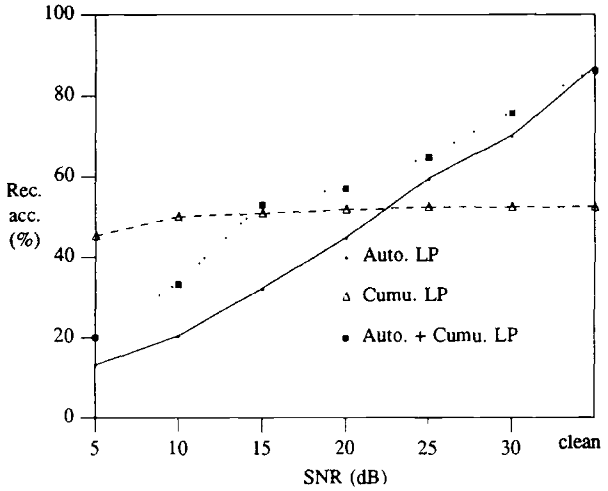

Fig. 6: Recognition accuracy for the alpha-digit recognition task as a function of SNR for additive white Gaussian noise.

4. SPEECH RECOGNITION EXPERIMENTS WITH ADDITIVE COLORED GAUSSIAN NOISE

In the preceding section, recognition performance was studied for speech corrupted by additive white Gaussian noise. In this section, we consider recognition of speech corrupted by the addition of colored Gaussian noise. Colored Gaussian noise was generated by filtering the white Gaussian noise by a second order AR filter (with coefficients -0.8018 and 0.3995). At 10 kHz sampling frequency, this colored noise has a resonance located at 1405 Hz with a bandwidth of 1460 Hz.

Figures 7 and 8 show the recognition accuracy as a function of SNR for the vowel recognition and the digit recognition tasks, respectively. Results are again shown for three different cases which utilize: 1) cepstral coefficients from the autocorrelation method, 2) cepstral coefficients from the cumulant-based methods, and 3) cepstral coefficients from the autocorrelation method + cepstral coefficients from the cumulant-based method. As can be seen, the results obtained for additive colored Gaussian noise are qualitatively similar to those observed earlier with additive white Gaussian noise. The fact that the method is insensitive to the spectrum of the disturbing noise, is noteworthy. We do not know of any other proposed method for which this property has been demonstrated.

5. CONCLUSIONS

In this paper we have proposed the use of cumulant-based LP analysis for speech recognition in the presence of noise. This method assumes the speech signal to be *non-Gaussian*. We have shown that cepstral coefficients derived by this method are quite insensitive to additive *Gaussian* noise which can be *white* or *colored*. We have compared the performance of a recognizer based on these estimates to one that uses LP estimates derived from the autocorrelation function. We find that at low SNR (below about 20 dB) the cumulant based estimates outperform the autocorrelation-based estimates.

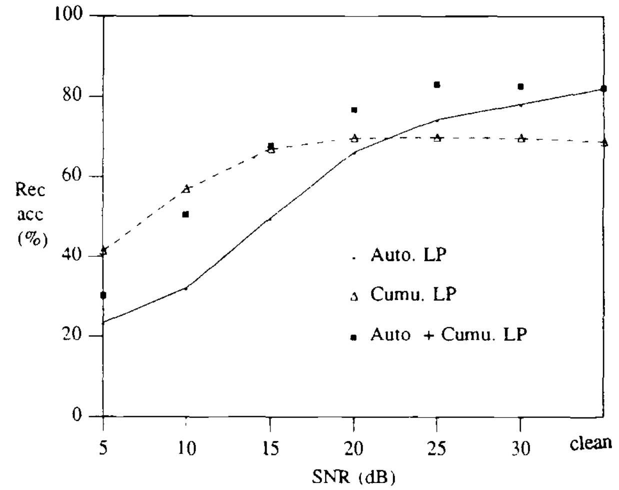

Fig. 7: Recognition accuracy for the vowel recognition task as a function of SNR for additive colored Gaussian noise.

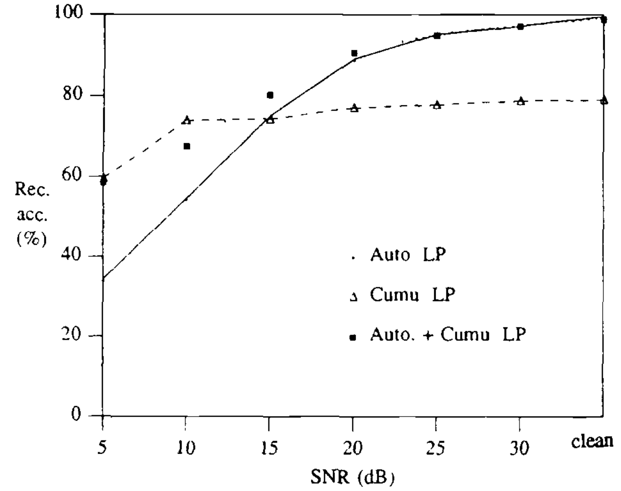

Fig. 8: Recognition accuracy for the digit recognition task as a function of SNR for additive colored Gaussian noise.

At higher SNRs the reverse is true. The reasons for this behavior are not yet understood. However, we have shown that by combining the two estimates, we can achieve recognition accuracy that is better than that of the conventional recognizer at all SNRs. Note that this improvement is obtained for both *white* and *colored* Gaussian noise, and without the need to first estimate the level or spectrum of the noise.

We have not yet made a detailed comparison of this method with other methods of combating noise in speech recognition. However, we can present one such comparison which is shown in Fig. 9. We see from this figure that our method compares favorably with the projection-based method of Mansour and Juang [5] for SNRs greater than about 20 dB, but is not as effective as that method for lower SNRs.

An important point which we are investigating at present, is the reason for the failure of the cumulant-based method for high SNRs. It is intuitively clear from the results shown in Figs. 4-8, that if the performance can be improved at high SNRs, then the method would provide a very robust speech recognition. Our conjecture is that improvements in the procedure for estimating cumulants might be the key to the problem.

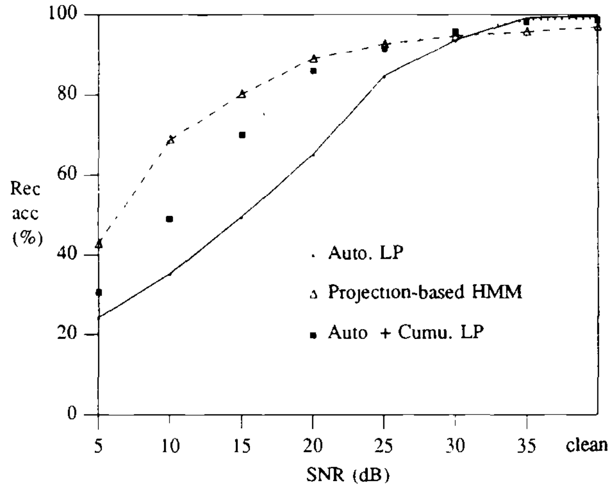

Fig. 9: Recognition accuracy for the digit recognition task as a function of SNR for additive white Gaussian noise.

References

[1] J. Makhoul, "Linear prediction: A tutorial review", Proc. IEEE, Vol. 63, pp. 561-580, 1975.

[2] Y. Ephraim, J.G. Wilpon and L.R. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments", Proc. ICASSP, pp. 1324-1327, 1987.

[3] O. Ghitza, "Robustness against noise: The role of timing synchrony measurement", Proc. ICASSP, pp. 2372-2375, 1987.

[4] D. Mansour and B.H. Juang, "The short-time modified coherence representation and noisy speech recognition", IEEE Trans. ASSP-37, pp. 795-804, 1989.

[5] D. Mansour and B.H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition", IEEE Trans. ASSP-37, pp. 1659-1671, 1989.

[6] F.K. Soong and M.M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise", IEEE Trans. ASSP-36, pp. 41-48, 1988.

[7] K.K. Paliwal, "Neural net classifiers for robust speech recognition under noisy environments", Proc. ICASSP, pp. 429-432, 1990.

[8] C.L. Nikias and M.R. Raghuveer, "Bispectrum estimation: A digital signal processing framework", Proc. IEEE, Vol. 75, pp. 869-891, 1987.

[9] M.R. Raghuveer and C.L. Nikias, "Bispectrum estimation: A parametric approach", IEEE Trans. ASSP-33, pp. 1213-1230, 1985.

[10] G.B. Giannakis and J.M. Mendel, "Identification of nonminimum phase systems using higher order statistics", IEEE Trans. ASSP-37, pp. 360-377, 1989.

[11] L.R. Rabiner, "A tutorial on Hidden Markov models and selected applications in speech recognition", Proc. IEEE, Vol. 77, pp. 257-286, 1989.

[12] B.H. Juang, L.R. Rabiner and J.G. Wilpon, "On the use of bandpass liftering in speech recognition", IEEE Trans. ASSP-35, pp. 947-954, 1987.