Statistical Properties of an LPC Distance Measure

JOSÉ M. TRIBOLET, MEMBER, IEEE, LAWRENCE R. RABINER, FELLOW, IEEE, AND MAN MOHAN SONDHI

Abstract—Several distance measures have been proposed for comparing sets of LPC coefficients. The most popular one has been the "log likelihood ratio" proposed by Itakura [1]. In this paper we discuss this measure (strictly speaking, a somewhat generalized version of it) from both a theoretical and a practical point of view. We derive its statistical properties both when the reference vector is known and when it is estimated from the data. We also show how these properties are affected by windowing, additive noise, and preemphasis. We present results of extensive simulations in support of the theoretical predictions. Finally, we argue that de Souza's [2] recent criticism of this measure is unjustified.

I. Introduction

Since a number of modern speech processing systems use linear prediction coefficients (LPC) as the basis of a speech representation [3]-[6], there has been a great deal of research on measures for comparing or computing the distance between sets of LPC coefficients. Based on the pioneering statistical analyses of Mann and Wald [7], Itakura [1] proposed a distance measure which he called the log likelihood ratio. This name is appropriate if the speech signal is modeled as the output of a linear system excited by Gaussian white noise. In Section II we will point out that it is possible to drop the requirement that the excitation be Gaussian. From this more general point of view, we will define a statistic which has a χ^2 distribution. Itakura's measure is just a monotonic function of this statistic. For want of a better term, however, we will continue calling it a log likelihood ratio.

This measure has been used in a variety of applications, including noise studies by Sambur and Jayant [8], LPC vocoder studies by Makhoul *et al.* [9], as a quality measure by Goodman *et al.* [10], and by Crochiere *et al.* [11].

The log likelihood ratio is attractive as a measure of distance for several reasons, including:

1) Its statistical properties are theoretically understood.

2) It can be efficiently computed using a single $(p+1)$ point dot product, where p is the order of the LPC analysis [1].

3) A physical interpretation can be attributed to the distance in terms of exciting the LPC model for one frame from the data of the other frame [12].

4) It can be interpreted in terms of spectral dissimilarity between the LPC spectra of the 2 frames [12].

Although alternative LPC distance measures have been proposed [12], the log likelihood ratio has remained the most popular for the reasons given above.

Manuscript received September 8, 1978; revised December 22, 1978.

J. M. Tribolet was with Bell Laboratories, Murray Hill, NJ 07094. He is now with the Department of Electrical Engineering, Instituto Superior Técnico, Lisbon, Portugal.

L. R. Rabiner and M. M. Sondhi are with Bell Laboratories, Murray Hill, NJ 07094.

In this paper we reexamine the statistical model that leads to the log likelihood ratio, and give experimental results which verify the validity of this model, both for the case when the reference LPC set is known exactly, and for the case when both the reference and test LPC sets are estimated from the data. We discuss the effects of the different LPC analysis methods, preemphasis, and additive noise on the statistical distributions. Using the results derived in this paper, we are able to give an explanation of de Souza's finding that there were significant discrepancies between predicted and experimentally observed distributions of the log likelihood measure. Finally, we discuss some practical implications of the results of our experiments for speech processing applications.

II. Statistical Properties of LPC Distances

The theoretical foundation for the statistical analysis of LPC distances was originally given by Mann and Wald [7]. In this section we review the relevant theory, present the key results of Mann and Wald using modern notation, and discuss its application to log likelihood ratios.

A. The Model

Consider an all-pole stable system of the form

$$y_n = -\sum_{k=1}^{p} a_k y_{n-k} + x_n \tag{1}$$

where the input samples x_n, $-\infty < n < \infty$ are statistically-independent identically-distributed random variables. We will assume that their distribution has mean zero and variance σ^2, and that it has finite higher moments. Stochastic difference equations of this type arise in many situations, e.g., as a model of a speech signal.

B. Parameter Estimation

Assuming that a given speech signal can be represented as the output of the system of (1), a standard problem is to estimate the parameters $a_1, \cdots, a_p$ and the variance σ^2 from just a knowledge of N output samples. A reasonable method of estimating these quantities is to use the minimum mean-squared error (MMSE) criterion. Thus consider the $N' = N - p$ equations obtained from (1) with $n = m, m-1, m-2, \cdots, m-N'+1$, respectively. (Note that knowledge of N samples allows one to write only N' equations.) These equations can be written in matrix notation as follows:

$$\boldsymbol{y} = -\boldsymbol{Y}\boldsymbol{a} + \boldsymbol{x} \tag{2}$$

where

$\boldsymbol{y}$ is a column vector with components $y_m, \cdots, y_{m-N'+1}$;

0096-3518/79/1000-0550$00.75 © 1979 IEEE

Y is a $N' \times p$ matrix whose components are

$$Y_{ij} = y_{m+1-i-j}, \quad i = 1, \cdots, N', \quad j = 1, \cdots, p; \tag{3}$$

$\boldsymbol{a}$ is a column vector with components $a_1, \cdots, a_p$;
$\boldsymbol{x}$ is a column vector with components $x_m, \cdots, x_{m-N'+1}$.

Form an estimate $\hat{\boldsymbol{y}}$, of $\boldsymbol{y}$ given by

$$\hat{\boldsymbol{y}} = -\boldsymbol{Y}\hat{\boldsymbol{a}} \tag{4}$$

where $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{a}}$ are defined analogously to $\boldsymbol{y}$ and $\boldsymbol{a}$. Then the MMSE estimate $\hat{\boldsymbol{a}}$ is obtained by minimizing

$$\begin{aligned} e &= (\boldsymbol{y} - \hat{\boldsymbol{y}})' (\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} + \boldsymbol{Y}\hat{\boldsymbol{a}})' (\boldsymbol{y} + \boldsymbol{Y}\hat{\boldsymbol{a}}) \end{aligned}$$

where "$'$" denotes matrix transposition. The usual minimization by setting the gradient equal to zero gives

$$\boldsymbol{Y}'\boldsymbol{Y}\hat{\boldsymbol{a}} = -\boldsymbol{Y}'\boldsymbol{y} \tag{5}$$

whose solution is the required estimate. Note that $\boldsymbol{Y}'\boldsymbol{Y}$ is just N' times the estimated $p \times p$ covariance matrix $\hat{\boldsymbol{\Phi}}$ of the output process $\boldsymbol{y}$. Thus (5) may be written as

$$N'\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{a}} = -\boldsymbol{Y}'\boldsymbol{y}. \tag{6}$$

Substituting for $\boldsymbol{y}$ from (2) gives

$$N'\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{a}} = N'\hat{\boldsymbol{\Phi}}\boldsymbol{a} - \boldsymbol{Y}'\boldsymbol{x}$$

or

$$N'\hat{\boldsymbol{\Phi}}\boldsymbol{\Delta} = -\boldsymbol{Y}'\boldsymbol{x} \tag{7}$$

where we have defined $\boldsymbol{\Delta} = \hat{\boldsymbol{a}} - \boldsymbol{a}$.

As is well known, (6) has a more convenient form in terms of the $(p + 1) \times (p + 1)$ covariance matrix and the $(p + 1)$ dimensional coefficient vector obtained from $\boldsymbol{a}$ by adding a component $a_0 = 1$. Denoting this augmented vector by $\boldsymbol{\alpha}$ and the $(p + 1) \times (p + 1)$ covariance matrix by $\boldsymbol{\Psi}$, (6) may be written as

$$\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}} = \hat{\sigma}^2 \boldsymbol{u}. \tag{8}$$

Here $\boldsymbol{u}$ is the unit vector with components $1, 0, \cdots, 0$; $\hat{\boldsymbol{\Psi}}$, $\hat{\boldsymbol{\alpha}}$, and $\hat{\sigma}$ are estimates of $\boldsymbol{\Psi}$, $\boldsymbol{\alpha}$, and σ, respectively, with

$$\hat{\sigma}^2 = \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}}. \tag{9}$$

C. The Estimation Error

With the above definitions the following theorem holds.

Theorem: In the limit as $N' \to \infty$, the components of the scaled LPC error vector $\sqrt{N'}\boldsymbol{\Delta}$ are jointly Gaussian. Their mean asymptotically approaches zero and their covariance matrix approaches Λ, such that

$$\Lambda = \lim_{N' \to \infty} \hat{\sigma}^2 \hat{\boldsymbol{\Phi}}^{-1} = \sigma^2 \boldsymbol{\Phi}^{-1}. \tag{10}$$

From the definition of a joint Gaussian distribution, any $q < p$ components are also jointly Gaussian with a covariance matrix obtained from Λ by selecting the respective rows and columns. Also, since $\hat{\boldsymbol{a}}$ differs from $\boldsymbol{\Delta}$ by the *fixed* vector $\boldsymbol{a}$, it follows that the components of $\hat{\boldsymbol{a}}$ are also Gaussian with the same covariance matrix.

A rigorous proof of this theorem was given by Mann and Wald [7] in a rather long and difficult paper. Their result is quoted often, yet we have not come across a simpler derivation. In the Appendix we give a heuristic proof which sacrifices some amount of rigor in order to gain some insight into the basic ideas involved.

D. Hypothesis Testing–Case 1: Reference LPC Vector Known

Suppose a vector $\hat{\boldsymbol{a}}$ has been estimated from N samples of a given signal $\boldsymbol{y}$ as above. It is often of interest (e.g., in word recognition tasks) to test the hypothesis that the signal was generated by (1) with a *specified* vector $\boldsymbol{a} = \boldsymbol{a}_0$. Call this hypothesis H_0. Then, defining $\boldsymbol{\Delta}_0 = \hat{\boldsymbol{a}} - \boldsymbol{a}_0$, we note that asymptotically under H_0 the vector $\sqrt{N'}\boldsymbol{\Delta}_0$ has jointly Gaussian components with covariance $\sigma^2\boldsymbol{\Phi}^{-1}$. From this it follows that under H_0 the quantity $N'\boldsymbol{\Delta}_0'\boldsymbol{\Phi}\boldsymbol{\Delta}_0/\sigma^2$ has asymptotically a χ^2 distribution with p degrees of freedom. [Note that if the specified vector $\boldsymbol{a}_0$ has $q < p$ components, then the distribution is still χ^2 with p degrees of freedom. This is because the specification is equivalent to specifying a p-component vector with $p - q$ trailing zeros.] Defining the statistic

$$l(\boldsymbol{a}_0, \hat{\boldsymbol{a}}) = \frac{N'}{\hat{\sigma}^2} \boldsymbol{\Delta}_0' \boldsymbol{\Phi} \boldsymbol{\Delta}_0, \tag{11}$$

in terms of estimated quantities, we therefore see that asymptotically $l(\boldsymbol{a}_0, \hat{\boldsymbol{a}})$ has a χ^2 distribution with p degrees of freedom. Equation (11) can be written in terms of $\hat{\boldsymbol{\Psi}}$ as

$$l(\boldsymbol{a}_0, \hat{\boldsymbol{a}}) = N' \frac{(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)'\hat{\boldsymbol{\Psi}}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)}{\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}}}. \tag{12}$$

Once we have a statistic with a χ_p^2 distribution, the hypothesis H_0 can be accepted or rejected by comparing it to prespecified thresholds.

E. "Log Likelihood Ratio"

Equation (12) can be put into a particularly simple form by using (8). Premultiplying this equation, in turn, by $\hat{\boldsymbol{\alpha}}'$ and $\boldsymbol{\alpha}_0'$ and remembering that $\boldsymbol{u}$ has components $1, 0, \cdots, 0$ we get

$$\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}} = \hat{\sigma}^2 \tag{13a}$$

$$\boldsymbol{\alpha}_0'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}} = \hat{\sigma}^2. \tag{13b}$$

Thus,

$$(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}} = 0. \tag{13c}$$

Then, it immediately follows from (12) that

$$l(\boldsymbol{a}_0, \hat{\boldsymbol{a}}) = N' \left[\frac{\boldsymbol{\alpha}_0'\hat{\boldsymbol{\Psi}}\boldsymbol{\alpha}_0}{\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}}} - 1 \right]. \tag{14}$$

The first term on the right-hand side of (14) is proportional to the likelihood ratio; therefore *the log likelihood ratio is a monotonic function of* $l(\boldsymbol{a}_0, \hat{\boldsymbol{a}})$. Clearly, any threshold for $l(\boldsymbol{a}_0, \hat{\boldsymbol{a}})$ corresponds to a unique threshold for the log likelihood ratio. Hence, for hypothesis testing the two are entirely equivalent. Thus if L denotes the log-likelihood ratio, then

$$L = N' \log_e \left[1 + \frac{l}{N'} \right]. \tag{15}$$

If $p_l(x)$ denotes the density function of l and $p_L(x)$ the density function of L, then

$$p_L(x) = e^{x/N'} p_l\,[N'(e^{x/N'} - 1)]. \tag{16}$$

On the scale to which the figures in this paper are drawn, the two functions are in fact indistinguishable. However, the point of view presented here is somewhat more appealing than the usual justification for the log likelihood ratio. (There is no need to assume Gaussian inputs or to approximate $\log_e \cdot (1 + x)$ by x as is usually done).

F. Hypothesis Testing–Case 2: Reference LPC Vector Estimated

In most applications the true vector $\boldsymbol{a}$ is not available for comparison. What is known instead is a MMSE estimate of $\boldsymbol{a}$, called the reference template $\hat{\boldsymbol{a}}_R$, obtained from some data $\boldsymbol{y}_R$. Let a test estimate $\hat{\boldsymbol{a}}_T$ be obtained from some data $\boldsymbol{y}_T$ independent of $\boldsymbol{y}_R$. We are then interested in testing the hypothesis H_0 that $\boldsymbol{y}_T$ and $\boldsymbol{y}_R$ were generated by the same underlying vector $\boldsymbol{a}$.

Assuming $\boldsymbol{y}_R$ and $\boldsymbol{y}_T$ each consists of N samples, we note that the components of the vector

$$\sqrt{N'}(\hat{\boldsymbol{a}}_T - \hat{\boldsymbol{a}}_R) = \sqrt{N'}(\hat{\boldsymbol{a}}_T - \boldsymbol{a}) - \sqrt{N'}(\hat{\boldsymbol{a}}_R - \boldsymbol{a}) \tag{17}$$

are again Gaussian under H_0. However, their covariance matrix is 2Λ because $\hat{\boldsymbol{a}}_T$ and $\hat{\boldsymbol{a}}_R$ are independent and identically distributed. Therefore, under H_0 the statistic

$$l(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{a}}_R) = \frac{N'}{2}\left[\frac{\hat{\boldsymbol{a}}_R'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{a}}_R}{\hat{\boldsymbol{a}}_T'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{a}}_T} - 1\right] \tag{18}$$

has a χ^2 distribution with p degrees of freedom.

G. Effect of Windowing and Preemphasis in the "Autocorrelation" Method

The derivations above have tacitly assumed the "covariance" method of LPC analysis. If the "autocorrelation" method is used, two additional factors must be taken into account—the effect of windowing and of preemphasis by a fixed (generally first-order) filter. We consider these effects in this section.

In the autocorrelation method the N given speech samples y_n are multiplied by a window w_n to give the windowed signal

$$\begin{aligned} v_n &= y_n w_n, \qquad 0 \leqslant n \leqslant N - 1 \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{19}$$

The windowed signal is then used to estimate $\boldsymbol{a}$ and σ^2 as before, (except that now the dimension of the vectors in (2) is N rather than N').

One effect of this windowing can best be described in the frequency domain. Thus, transforming (19) gives

$$\begin{aligned} V(j\omega) &= Y(j\omega) * W(j\omega) \\ &= [X(j\omega) \cdot H(j\omega)] * W(j\omega) \end{aligned} \tag{20}$$

where "$*$" denotes convolution and $H(j\omega)$ is the transfer function (i.e., the Fourier transform of the impulse response) of the system of (1). If $|H(j\omega)|$ (and hence $|X(j\omega)H(j\omega)|$) has strong peaks, then $V(j\omega)$ can be quite different from $Y(j\omega)$ unless the window is tapered and very long compared to the effective length of the impulse response. We have not been able to quantify this effect. However, simulations indicate that the error introduced is such that even for Case 2 estimates, the distribution of $\hat{\boldsymbol{\alpha}}_T - \hat{\boldsymbol{\alpha}}_R$ is different from χ_p^2, unless the window is quite long. (The data of Fig. 6 in Section III illustrates this point.)

One way to partially alleviate this problem is to preemphasize the speech signal before windowing with the aim of reducing the *dynamic* range of $|X(j\omega)H(j\omega)|$, or equivalently reducing the effective duration of the impulse response of the system. The "smearing" produced by the convolution of (20) would then be less severe. For speech signals, a first-order preemphasis filter, (e.g., the one given by (29)] can make a significant improvement in the accuracy of autocorrelation analysis. Note that since there is no windowing in the covariance method, there is no need for preemphasis. If used, it just increases the order of the system (by 1 for a first-order preemphasis).

If one can assume that the window is *long compared to the impulse response*, then the effect of windowing can be taken into account by merely replacing N' in the above analysis by an effective number of samples N_{eff}. To estimate N_{eff}, let us multiply both sides of (2) by a diagonal matrix $\boldsymbol{W}$, whose diagonal elements are the window weights. Then for $N \gg p$, and a slowly varying window, (2) becomes

$$\begin{aligned} \boldsymbol{W}\boldsymbol{y} = \boldsymbol{v} &= - \boldsymbol{W}\boldsymbol{Y}\boldsymbol{a} + \boldsymbol{W}\boldsymbol{x} \\ &\approx - \boldsymbol{V}\boldsymbol{a} + \boldsymbol{W}\boldsymbol{x}. \end{aligned} \tag{21}$$

Here $\boldsymbol{v}$ is the weighted output vector and $\boldsymbol{V}$ is obtained from $\boldsymbol{v}$ exactly as $\boldsymbol{Y}$ was obtained from $\boldsymbol{y}$. (Of course the long dimension in (21) is understood to be N.) Thus the situation is as before with $\boldsymbol{x}$ replaced by $\boldsymbol{W}\boldsymbol{x}$ and $\boldsymbol{Y}$ by $\boldsymbol{V}$. Let w_2 and w_4 represent the averages of the second and fourth power of the window function, i.e.,

$$w_2 = \frac{1}{N}\sum_{i=0}^{N-1} w^2(i)$$

$$w_4 = \frac{1}{N}\sum_{i=0}^{N-1} w^4(i).$$

Then it is seen that

$$\begin{aligned} \boldsymbol{V}'\boldsymbol{V} &\approx w_2 \boldsymbol{Y}'\boldsymbol{Y} \\ \boldsymbol{V}'\boldsymbol{W}\boldsymbol{x} &\approx w_2 \boldsymbol{Y}'\boldsymbol{x} \\ \boldsymbol{V}'\boldsymbol{W}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{W}\boldsymbol{V}' &\approx w_4 \boldsymbol{Y}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{Y}. \end{aligned}$$

Substituting these values in the appropriate equations in the Appendix, it can be seen that, to the extent that these approximations are valid, there is no bias for the autocorrelation method. However, the effective number of samples is

$$\begin{aligned} N_{eff} &\approx \frac{w_2^2}{w_4} N \\ &\triangleq \beta N. \end{aligned} \tag{22}$$

Here $\beta = 1$ for a rectangular window and $\beta = 0.55$ for a Hamming window. (We believe 0.55 to be much more accurate than the value 0.3975 suggested by Sambur and Jayant [8].) For consistency we may define

$$N_{eff} = N - p = N' \tag{23}$$

for the covariance method.

H. The Effects of Additive White Noise

The effects of additive noise on LPC analysis have been studied by Sambur and Jayant [9], Lim and Oppenheim [13], and Boll [14]. Additive noise degrades the LPC analysis by introducing distortions of the signal spectrum in the valleys. In statistical terms, the estimation procedure of Section II-B gives a biased estimate of $\boldsymbol{a}$. Thus suppose instead of y, one is given the vector

$$z = y + \boldsymbol{n} \tag{24}$$

where $\boldsymbol{n}$ is a statistically independent white noise. Then if Z is obtained from z as Y was from y, the equation for the estimate $\boldsymbol{a}$ becomes

$$Z'Z\hat{a} = -Z'z. \tag{25}$$

An analysis similar to the noiseless case shows that asymptotically as N' becomes large, the mean of $\sqrt{N'}\,\hat{a}$ is given by

$$E[\sqrt{N'}\,\hat{a}] = \sqrt{N'}\boldsymbol{\Phi}_z^{-1}\boldsymbol{\Phi}\boldsymbol{a} \tag{26}$$

where

$$\boldsymbol{\Phi}_z = \frac{1}{N'}E[Z'Z] \tag{27}$$

and $\boldsymbol{\Phi}$ is defined in (10).

The components of $\sqrt{N'}\,\hat{a}$ again become Gaussian as $N \to \infty$; their means are given by (26) and their covariance matrix asymptotically approaches Λ_z given by

$$\Lambda_z = (\sigma^2 + \sigma_n^2)\,\boldsymbol{\Phi}_z^{-1} \tag{28}$$

where σ_n^2 is the variance of the added noise. Since y is not observable, the mean of (26) cannot be computed, and therefore the distribution of $\sqrt{N'}(\hat{a} - a)$ cannot be computed. The distribution obtained by assuming the mean to be zero gives a very poor fit to measurements, as we shall see in Section III-C.

Consider now the case when a test estimate $\hat{a}_T$ and a reference template $\hat{a}_R$ have been obtained in the presence of noise *with the same statistical properties*. In this case the distribution of $\hat{a}_T - \hat{a}_R$ is much less affected by the noise. Note that $\sqrt{N'}\,(\hat{a}_T - \hat{a}_R)$ has *zero mean* and a covariance given by $2\Lambda_z$. Now in (28) the matrix $\boldsymbol{\Phi}_z$ can be estimated from z. Thus if $\sigma^2 + \sigma_n^2$ could be estimated, we would have an estimate of Λ_z. If we define $\boldsymbol{\Psi}_z$ analogously to $\boldsymbol{\Psi}$, then as for the noiseless case we can use $\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\Psi}}_z\hat{\boldsymbol{\alpha}}$ as an estimate of $\sigma^2 + \sigma_n^2$. This estimate can be shown to be biased. Nevertheless, it allows us to define a statistic $l(\hat{a}_T, \hat{a}_R)$ exactly as in (18) for the noiseless case, with $\hat{\boldsymbol{\Psi}}$ replaced by $\hat{\boldsymbol{\Psi}}_z$. As we will see in Section III-C, the bias in the estimate of $\sigma^2 + \sigma_n^2$ is not very large.

In summary, we have shown that the estimate $\hat{a}$ is strongly affected by additive noise. For Case 1 estimates, the distribution is very different from χ_p^2. However, for Case 2 estimates the bias in the mean cancels, and the estimate of (18) appears to be quite insensitive to the error in the estimate of variance.

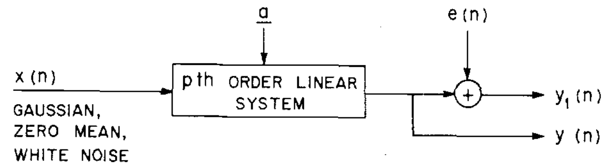

Fig. 1. Block diagram of system used to investigate log likelihood ratio distances for LPC coefficients.

III. Experimental Results

To test the validity of the analysis equations of Section II, a Gaussian zero-mean white-noise signal was used to excite a linear system of the type shown in Fig. 1. In order to verify that the distribution of the input is unimportant *one* test condition (see Fig. 4) was generated with independent uniform noise samples as the excitation. The coefficients of the linear system (the vector $\boldsymbol{a}$) were determined from an LPC analysis of a steady-state portion of the vowel *a* as in *fa*ther from a section of natural voiced speech. The autocorrelation method of LPC analysis was used. Two sets of LPC coefficients were obtained; one set $\boldsymbol{a} = \boldsymbol{a}_1$, was from a 10 pole analysis ($p = 10$) of the vowel; the other set $\boldsymbol{a} = \boldsymbol{a}_2$ was from a 10 pole analysis of the preemphasized vowel, where the preemphasis network was of the form

$$H_p(z) = 1 - 0.95z^{-1}. \tag{29}$$

Fig. 2 shows plots of the acoustic waveform of the original vowel [Fig. 2(a)], the signal spectrum and the LPC spectrum of the set $\boldsymbol{a} = \boldsymbol{a}_1$, [Fig. 2(b)], the preemphasized acoustic waveform [Fig. 2(c)], and the preemphasized signal spectrum and the resulting LPC spectrum of the set $\boldsymbol{a} = \boldsymbol{a}_2$ [Fig. 2(d)]. It is readily seen that preemphasis reduces the dynamic range of the signal spectrum (and the LPC spectrum) by about 10 dB. Table I gives the actual coefficients for the 2 sets $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$.

For some of the tests a zero-mean white Gaussian noise e_n was added to the output y_n at signal-to-noise ratio S/N where S/N was either 20 dB or 10 dB.

A total of 1000 independent frames, each of duration 300 samples, were created using the system of Fig. 1 for each of the 4 input conditions, namely:

1) No preemphasis ($\boldsymbol{a} = \boldsymbol{a}_1$); no noise–$y(n)$ output
2) Preemphasis ($\boldsymbol{a} = \boldsymbol{a}_2$); no noise–$y(n)$ output
3) No preemphasis ($\boldsymbol{a} = \boldsymbol{a}_1$); noise added–$y_1(n)$ output
4) Preemphasis ($\boldsymbol{a} = \boldsymbol{a}_2$); noise added–$y_1(n)$ output.

For each of the 4 input conditions, 2 types of LPC analysis were performed. These were:

1) Covariance analysis with a frame length of 300 samples, and a number of poles $p = 10$.

2) Autocorrelation analysis with a frame length of 300 samples, and a number of poles $p = 10$. Both a rectangular window and a Hamming window were used.

In addition, for each of the combinations of inputs and analyses, two measurements of the log likelihood ratio were made. These were:

1) Case 1 [(14)]–where the LPC vector $\boldsymbol{a}$ was known and only the test LPC vector was computed from the data.

2) Case 2 [(18)]–where both the reference LPC vector $\boldsymbol{a}_R$ and the test LPC vector $\boldsymbol{a}_T$ were computed from the data.

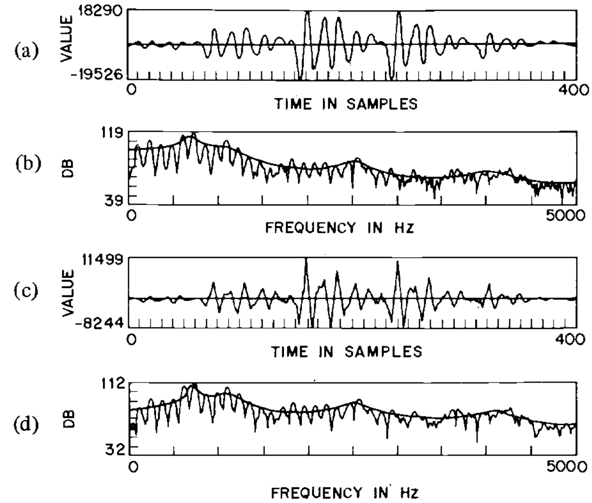

Fig. 2. Plots of the acoustic waveforms and the original and LPC log spectra for the 400 sample section of the vowel /a/ both without preemphasis [(a) and (b)], and with preemphasis [(c) and (d)]. The LPC analysis was performed using the autocorrelation method with a Hamming window.

TABLE I
LPC Coefficients for the Original ($\boldsymbol{a}_1$) and Preemphasized ($\boldsymbol{a}_2$) Data

	$\underline{a}_1$	$\underline{a}_2$
0	1.0	1.0
1	-2.16800	-1.27323
2	1.85800	0.811009
3	-0.854807	-0.401772
4	0.510366	0.477587
5	-0.389642	-0.190580
6	0.700694	0.703581
7	-1.21652	-0.791582
8	0.896169	0.451886
9	-0.0942807	-0.102455
10	-0.118126	0.216100

In these cases the reference and test vectors were estimated from independent frames of data.

In the remainder of this section we present results on the different conditions described above and relate these results to the theoretical discussion of Section II.

A. *Case 1–No Additive Noise*

The first set of results given in Fig. 3 is for Case 1 when the LPC vector $\boldsymbol{a}$ is known. Fig. 3 shows plots of the theoretical and measured histograms of the log likelihood ratio for the following cases:

1) Covariance analysis–no preemphasis [Fig. 3(a)].

2) Autocorrelation analysis–rectangular window, no preemphasis [Fig. 3(b)].

3) Autocorrelation analysis–Hamming window, no preemphasis [Fig. 3(c)]. [β of (21) set to 0.55].

4) Covariance analysis–preemphasized [Fig. 3(d)].

5) Autocorrelation analysis–rectangular window, preemphasized [Fig. 3(e)].

6) Autocorrelation analysis–Hamming window, preemphasized [Fig. 3(f)]. [β of (21) set to 0.55].

A total of 998 measurements of the log likelihood ratio distance were used in all measured histograms. The data were obtained by using the first 998 nonoverlapping frames of output of the system of Fig. 1.

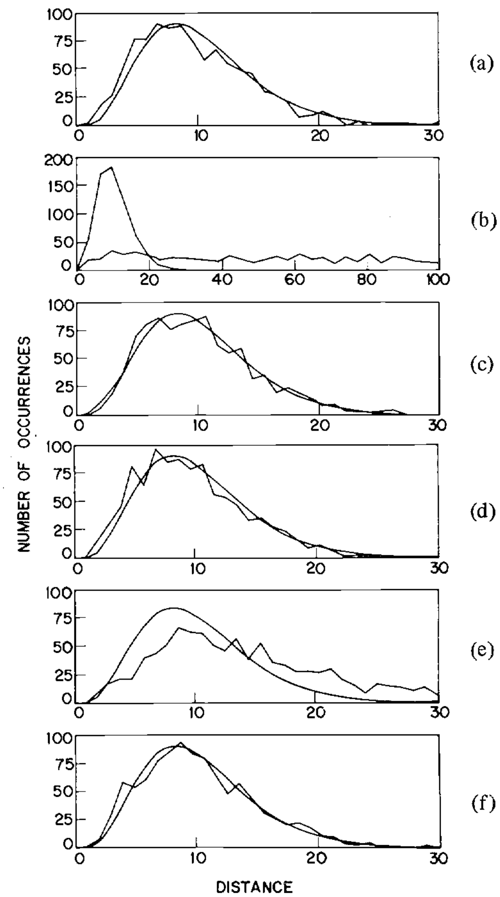

Fig. 3. Plots of the theoretical and measured histograms of the log likelihood ratio distance for the Case 1 computation without additive noise. Parts (a), (b), and (c) are for data without preemphasis, and parts (d), (e), and (f) are for the preemphasized data. Parts (a) and (d) are for the covariance analysis method; parts (b) and (e) are for the autocorrelation method with a rectangular window; parts (c) and (f) are for the autocorrelation method with a Hamming window. [Note the scale change in part (b).]

The extremely poor fit between the measured distribution and theoretical χ_p^2 distribution shown in Fig. 3(b) is anticipated by the discussion of Section II-G because of the poor sidelobes of a rectangular window. Fig. 3(e) clearly shows the improvement obtained by first preemphasizing the signal and then performing the LPC analysis.

In Fig. 4 we show plots of the distribution obtained when uniform noise was used as the excitation in Fig. 1 for cases corresponding to parts (a), (b), and (c) of Fig. 3. Clearly, the fits to the distributions are equally good for the uniform noise excitation as for the Gaussian noise excitation.

From Figs. 3 and 4 the following conclusions can be drawn.

1) There is excellent agreement between the measured distribution and the theoretical χ^2 distribution with $p = 10$ degrees of freedom for 4 of the 6 conditions.

2) The effect of preemphasis is to considerably improve the fit for the case of using a rectangular window with the autocorrelation method.

3) An effective length of $\beta = 0.55$ provides good fits for the Hamming window examples.

4) The distribution of the LPC estimates is quite insensitive

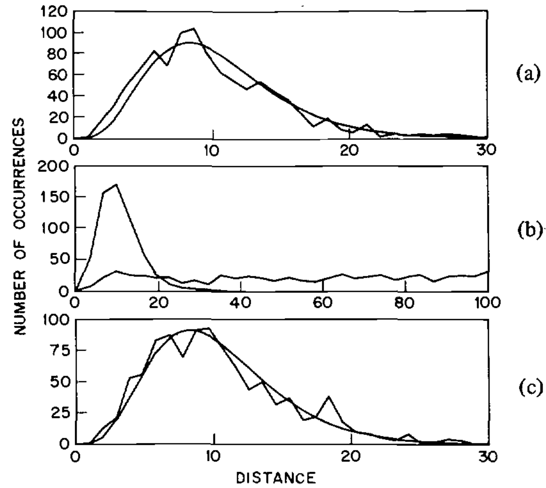

Fig. 4. Plots of the theoretical and measured histograms of the log likelihood ratio distance for Case 1 distance computation using a uniform noise excitation instead of a Gaussian noise. Parts (a)-(c) correspond to those of parts (a)-(c) of Fig. 3.

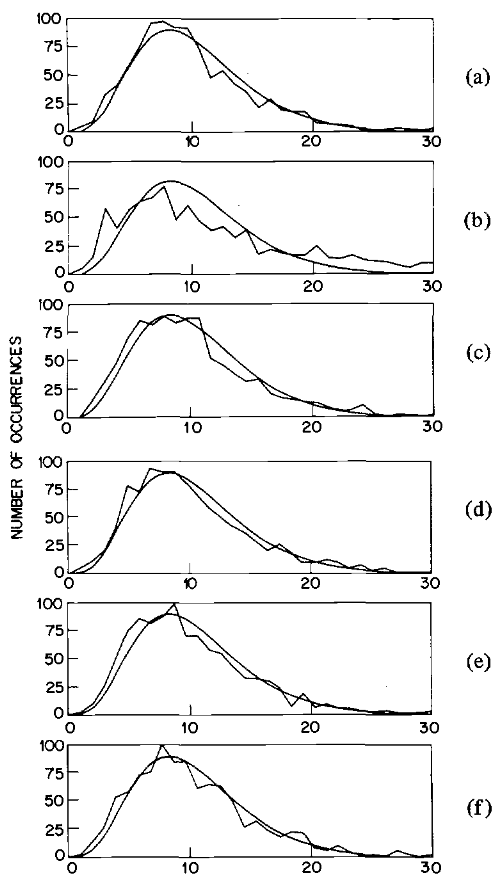

Fig. 5. Plots of the theoretical and measured histograms of the log likelihood ratio distance for the Case 2 distance computation without additive noise. Parts (a)-(f) correspond to those of Fig. 3.

to the distribution of the input exciting the linear system of (1).

B. Case 2–No Additive Noise

Fig. 5 shows plots of the measured and theoretical distributions of the log likelihood ratio for the Case 2 analysis methods

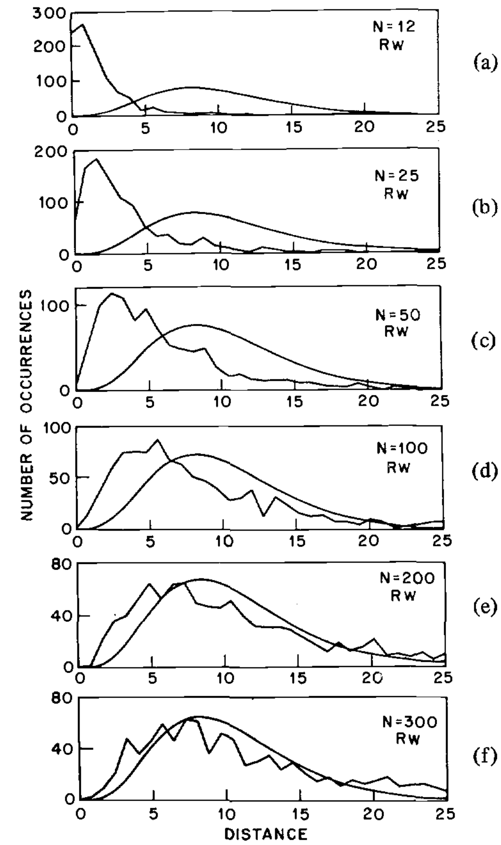

Fig. 6. Plots of the theoretical and measured histograms of the log likelihood ratio distance for the autocorrelation method using the Case 2 distance computation with an N point rectangular window. Results are given for several values of N.

in which both the reference and test LPC vectors were estimated from the data, and when no additive noise was used. The six plots are for the same 6 cases as shown in Fig. 3. It can be seen that the agreement between the measured and theoretical distributions of the log likelihood ratio is extremely good for all cases except the unpreemphasized autocorrelation analysis using the rectangular window where the agreement is somewhat worse than for the other cases. These examples (i.e., the data of Figs. 3, 4, and 5) essentially completely validate the statistical model of Section II.

To illustrate the effect of windowing on the distributions, Fig. 6 shows a series of plots of theoretical and measured histograms of the log likelihood ratio for the autocorrelation method with an N point rectangular analysis window. In these plots N varies from 12 to 300. The effective impulse response duration of the system was about 50 samples. We see from these plots that until N is about 200 or 300 samples (i.e., from 4 to 6 times the effective impulse response duration), the fits between the measured and theoretical distributions are rather poor.

C. Additive Noise Examples

To investigate the effects of additive zero-mean white Gaussian noise on the agreement between the theoretical and actual distributions of the log likelihood ratio, noise was added to the output signal $y(n)$ in Fig. 1 at signal-to-noise ratios of 10 dB and 20 dB. In Figs. 7 and 8 are shown plots of measured

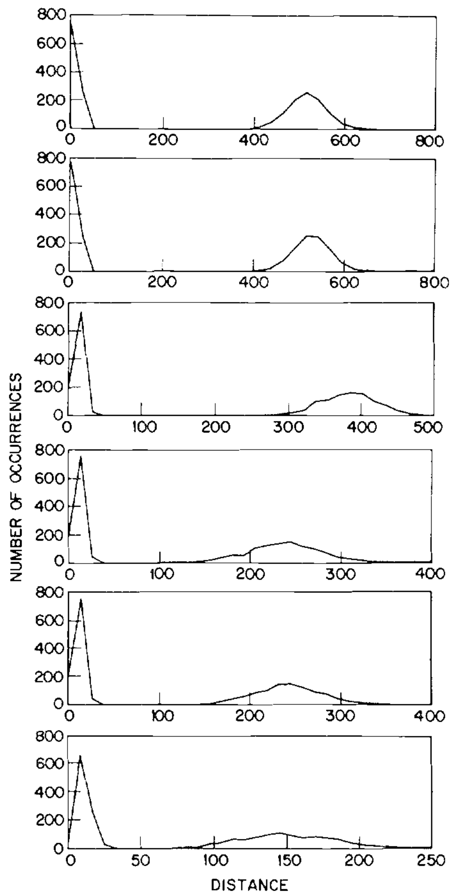

Fig. 7. Plots of the theoretical and measured histograms of the log likelihood ratio distance for the Case 1 distance computation with 10 dB signal-to-noise ratio additive noise. Plots (a)-(f) correspond to those of Fig. 3.

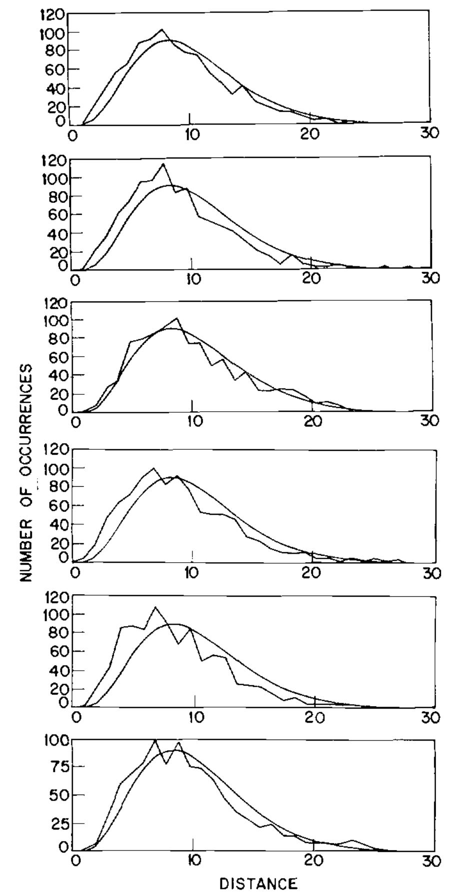

Fig. 8. Plots of the theoretical and measured histograms of the log likelihood ratio distance for the Case 2 distance computation with 10 dB signal-to-noise ratio additive noise. Plots (a)-(f) correspond to those of Fig. 3.

and theoretical distributions of the log likelihood ratio for both Case 1 (Fig. 7) and Case 2 (Fig. 8) for the 10 dB signal-to-noise ratio examples. (Essentially, equivalent results were obtained for the 20 dB cases).

From Fig. 7 it is seen that there is essentially no agreement between the theoretical and measured distributions for the Case 1 data since the estimate of the LPC set $\hat{a}$ from the noisy data was greatly in error, as discussed in Section II. However, as seen in Fig. 8, when one used the Case 2 method of estimating both reference and test LPC sets from the noisy data, the theoretical and measured distributions of the log likelihood ratio were essentially the same. Thus the error in the estimation of $\sigma^2 + \sigma_n^2$ mentioned in Section II-H is not significant.

D. Explanation of de Souza's Results

In addition to the sets of data discussed in Section III, the 25th order system used by de Souza was simulated with the system of Fig. 1. The LPC coefficients were identically those used by de Souza. Fig. 9(a) shows the frequency response, and Fig. 9(b) shows the impulse response of the linear system that was used. Several striking aspects of this system can be noted in Fig. 9. First we see that although a 25th order system was used, the first pole is of narrow bandwidth and low-center frequency, whereas the remaining poles are much higher in frequency. Due to the narrowness of the bandwidth of the lowest pole, the amplitude of the log spectrum is down on the order of 40 dB or more for the higher poles. Thus this linear system, although technically a 25th order system, could be well modeled as a second-order system. From Fig. 9(b) we see the result of the narrow bandwidth of the first pole is that the impulse response lasts for more than 1000 samples. Thus, for the autocorrelation method, it is necessary to use section lengths N greater than 1000.

De Souza compares the χ_{25}^2 distribution to one obtained experimentally from the output of the above 25th order system and demonstrates a very poor match. However, it is to be noted that 1) he used the autocorrelation method with rectangular 200 sample windows (from the results shown in Fig. 6 this is clearly inadequate); 2) he used Case 2 estimates $(\hat{a}_T - \hat{a}_R)$, but did not divide $l(\hat{a}_T, \hat{a}_T)$ by 2 as proven necessary in Section II-F. We believe that with the factor 0.5 included, and with tapered (e.g., Hamming) windows longer than 1000 samples, he would have obtained a much better match. De Souza does not show a comparison with an experimental distribution of estimates by the covariance method but states merely that the fit was "even worse." Presumably, the factor 0.5 would have significantly improved that fit also.

In support of these assertions Fig. 10 shows a plot of the measured and theoretical distributions of the log likelihood ratio obtained using the Case 1 estimate (a known) for the

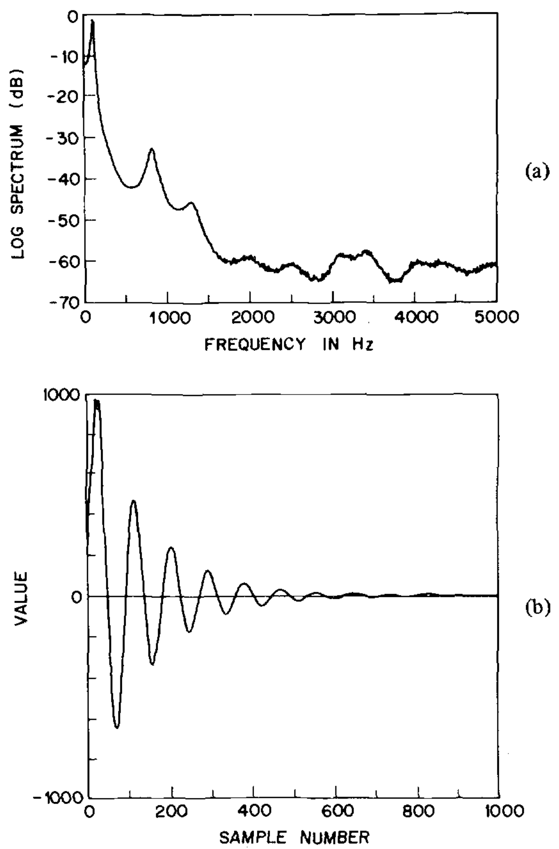

Fig. 9. The log spectrum in dB and the resulting impulse response of the 25th order system investigated by de Souza.

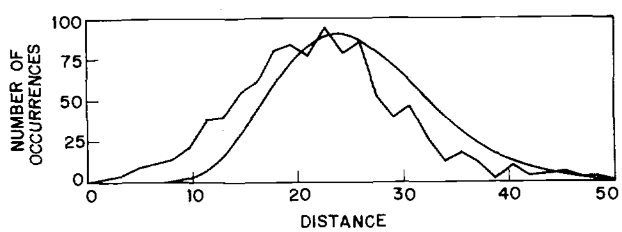

Fig. 10. Plots of the theoretical and measured histograms of the log likelihood ratio distance for the 25th order system of Fig. 8 using the Case 1 distance computation and the covariance analysis method.

covariance method with preemphasis of the data. The measured data do have a slightly smaller mean and variance than the theoretical χ^2 distribution for 25 degrees of freedom. However, the match is not nearly as bad as that suggested by de Souza.

IV. Application of Statistical Results to Speech Examples

We have shown that in the case of random inputs exciting linear systems, the measured properties of the log likelihood ratio agree closely with those predicted theoretically—namely, that the ratio for p-dimensional LPC vectors is χ^2 distributed with p degrees of freedom, provided p is at least equal to the order of the linear system. The remaining key question is the applicability of this result to actual speech signals.

For fricative sounds, the model studied here applied directly, and the distributions are as predicted. For voiced speech sounds, however, the measured distributions are *not* χ^2 distributed for any of the alternative cases we have discussed in this paper. This is because the assumptions used to derive the distribution break down for voiced speech sounds. For such sounds there is a random component of the excitation (e.g., modeling error, the high-frequency portion of many voiced sounds, etc.) which may plausibly have the properties assumed above. However, a large part of the energy in the excitation is quasi-periodic, and cannot be assumed to consist of statistically independent random samples. The effect of this component is to add a bias to the estimates and, of course, make the estimate of σ^2 larger than the variance of the random component. Thus knowledge of the distribution of the log likelihood ratio for random inputs does not solve the problem of providing thresholds in the case of voiced sounds. Nevertheless, word recognition algorithms based on the likelihood ratio are highly successful in practice [8], [9], [15]. For this we have the following plausible, but far from adequate, explanation.

Note that in a word recognition task what is of interest is the *sum* of the distances between many pairs (typically, 20 to 30) of LPC vectors. And the vectors are not all estimates of the same speech sound but, typically, of 5 or 6 different speech sounds. We suggest that the bias term becomes negligible when averaged over many different voiced sounds. In that case the total distance would still be approximately a sum of χ^2 distributions, except for a scaling of $\hat{\sigma}^2$. The exact scaling error is of course unknown, but it is plausible that a compromise threshold can be experimentally determined.

V. Summary

In this paper we have shown that the log likelihood ratio for p-dimensional LPC estimates is both theoretically and in practice χ^2 distributed with p degrees of freedom, provided p is at least equal to the order of the linear system which generated the data being analyzed. We have examined the effects of preemphasis, different LPC methods, different windows, and additive random noise on the measured distributions. Finally, we have given a plausible explanation why such a statistical model can be "roughly" applied to actual speech signals.

Appendix

A Heuristic Justification of Mann and Wald's Theorem (Section II-C)

In this Appendix we will derive the result of Mann and Wald on the distribution of the scaled error vector $\Delta\sqrt{N'}$. However, in order not to obscure the main arguments, we will assume the existence of various limiting distributions. (A rigorous proof of the existence of these limits is not trivial, and we refer the interested reader to [7] for the details.)

Our starting point is (7) of the text where we first introduced the error vector $\boldsymbol{\Delta}$. Let us rewrite that equation as

$$\hat{\boldsymbol{\Phi}}(\boldsymbol{\Delta}\sqrt{N'}) = -\frac{1}{\sqrt{N'}}\, Y'\boldsymbol{x}. \tag{A-1}$$

By assumption, (1) of the text represents a stable system; hence, its impulse response eventually goes to zero. Suppose that the number of samples $N = qM + p$, where M is larger than the effective length of the impulse response, and q is some large integer. Recalling that Y' has p rows and $N' = N - p$ columns and the vector x is N'-dimensional, we can partition Y and x into q pieces of length M along the N' dimension. Thus,

$$Y'x = [Y_1' \; Y_2' \cdots Y_q'] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} = \sum_{k=1}^{q} Y_k' x_k. \tag{A-2}$$

Now each of the q terms in this summation is a (zero-mean) random vector of dimension p. These q vectors are identically distributed. Further, as $M \to \infty$ (and the length of the impulse response becomes a vanishingly small fraction of M), it is clear that these vectors also become statistically independent. By the central limit theorem, therefore, the components become jointly Gaussian in the limit.

Now as N becomes large $\hat{\mathbf{\Phi}} \to \mathbf{\Phi}$. Since y is the output of a stable system, $\mathbf{\Phi}^{-1}$ exists. Then premultiplying both sides of (A-1) by $\mathbf{\Phi}^{-1}$ shows that in the limit as $N \to \infty$ the components of $\mathbf{\Delta}\sqrt{N'}$ are linear combinations of Gaussian variables, and therefore are themselves jointly Gaussian.

It remains to find the covariance matrix of $\mathbf{\Delta}\sqrt{N'}$. Denote this matrix by Λ as in the text. Then we have from (A-1)

$$E[\hat{\mathbf{\Phi}}(\mathbf{\Delta}\sqrt{N'})(\mathbf{\Delta}\sqrt{N'})'\hat{\mathbf{\Phi}}'] = \frac{1}{N'} E[Y'xx'Y] \triangleq \frac{1}{N'} E[Q] \tag{A-3}$$

where E denotes expectation and Q is the matrix on the previous line. We will presently show that

$$\frac{1}{N'} E[Q] = \sigma^2 \mathbf{\Phi}. \tag{A-4}$$

From (A-3) and (A-4), since $\hat{\mathbf{\Phi}} \to \mathbf{\Phi}$ and $\mathbf{\Phi}' = \mathbf{\Phi}$, it follows that

$$\Lambda = \sigma^2 \mathbf{\Phi}^{-1}, \tag{A-5}$$

which proves the theorem stated in the text.

To show that (A-4) holds, we write the rsth component of the matrix Q. From the definition of Y given in (3) of the text

$$Q_{rs} = \sum_{j,k=1}^{N'} y_{m+1-r-j} y_{m+1-s-k} x_{m+1-j} x_{m+1-k}. \tag{A-6}$$

Note that x_m is statistically independent of y_n for all indices m, n such that $m > n$. From this it follows that a term with $j > k$ in (A-6) has zero expectation. This is because in that case x_{m+1-k} is independent of the three quantities it multiplies and therefore its expectation (which is zero) factors out. By symmetry, a term with $k > j$ also has zero expectation. Therefore, only terms with $k = j$ remain. In each of these terms the y's are again independent of the x's. Therefore, the expectation of each term factors into the expectation of the y's and the expectation of the x's. Thus,

$$\begin{aligned} E[Q_{rs}] &= \sum_{k=1}^{N'} E[y_{m+1-r-k} y_{m+1-s-k}] E[x_{m+1-k}]^2 \\ &= \sigma^2 E\left[\sum_{k=1}^{N'} y_{m+1-r-k} y_{m+1-s-k}\right] \\ &= \sigma^2 E[(Y'Y)_{rs}] \\ &= N' \sigma^2 \mathbf{\Phi}_{rs} \end{aligned} \tag{A-7}$$

which proves (A-4).

Acknowledgment

The authors would like to thank Dr. C. Mallows for suggesting the method for deriving the distribution of the LPC parameters discussed in the Appendix, and Dr. V. Nair for helpful comments and criticisms of an early version of the manuscript.

References

[1] F. I. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol ASSP-23, pp. 67–72, Feb. 1975.

[2] P. V. de Souza, "Statistical tests and distance measures for LPC coefficients," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-25, pp. 554–559, Dec. 1977.

[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978, ch. 8.

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

[5] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[6] J. R. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[7] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica*, vol. 11, pp. 173–220, July–Oct. 1943.

[8] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 488–494, Dec. 1976.

[9] J. R. Makhoul, I. Viswanathan, L. Cosel, and W. Russel, "Natural communication with computers: Speech compression research at BBN," Bolt, Beranek, and Newman, Cambridge, MA, BBN Rep. 2976, Dec. 1974.

[10] D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman, "Objective and subjective performance of tandem connections of waveform coders with an LPC vocoder," *Bell Syst. Tech. J.*, vol. 58, pp. 601–629, Mar. 1979.

[11] R. E. Crochiere, L. R. Rabiner, N. S. Jayant, and J. M. Tribolet, "A study of objective measures for speech waveform coders," presented at 1978 Int. Zurich Seminar on Digital Commun., Mar. 1978.

[12] A. H. Gray, Jr., and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380–391, Oct. 1976.

[13] 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.

[14] S. F. Boll, "Suppression of noise in speech using the Saber method," in *Proc. 1978 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, OK, Apr. 1978, pp. 606–609.

[15] M. R. Sambur and L. R. Rabiner, "A statistical decision approach to the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 550–558, Dec. 1976.