# A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis

AUGUSTINE H. GRAY, JR., MEMBER, IEEE, AND JOHN D. MARKEL, MEMBER, IEEE

*Abstract*—**The purpose of this paper is to introduce a spectral-flatness measure into the study of linear prediction analysis of speech. A spectral-flatness measure is introduced to give a quantitative measure of "whiteness," of a spectrum. It is shown that maximizing the spectral flatness of an inverse filter output or linear predictor error is equivalent to the autocorrelation method of linear prediction. Theoretical properties of the flatness measure are derived, and compared with experimental results. It is shown that possible ill-conditioning of the analysis problem is directly related to the spectral-flatness measure and that prewhitening by a simple first-order linear predictor to increase spectral flatness can greatly reduce the amount of ill-conditioning.**

## I. INTRODUCTION

IN LINEAR prediction analysis of speech [1], a set of coefficients is obtained which can be used to define an all-zero digital filter. It has been noted that if the signal from which the coefficients are obtained is passed through this filter, the output will tend to be a whitened version of the input [2].

In this paper, the whitening process of this filter (referred to as an "inverse filter") is studied in some detail. First, a spectral-flatness measure is introduced for quantifying the "flatness" of any speech spectrum.

This numerical measure is shown to be physically meaningful by presenting several representative analysis conditions for two important classes of speech sounds, the voiced and unvoiced sounds.

Spectral–flatness measures for driving function models (models of the inverse filter output) of experimental results are presented to show that the theoretical predictions are accurate.

Next, relationships between the spectral-flatness measure and the possible ill-conditioned nature of the solution process for obtaining the inverse filter coefficients are discussed. It is shown that the spectral–flatness measure itself is a quantitative measure of ill-conditioning that can be used to explain several "rules of thumb" relating to necessary accuracy in computing inverse filters.

Finally, it is shown that preemphasis of the speech data by a first-order filter that maximizes the spectral flatness of its output is an important procedure for minimizing the ill-conditioning of the solution process.

Manuscript received December 12, 1972; revised September 9, 1973 and January 9, 1974. This research was supported under ONR Contract N00014-67-0118.

A. H. Gray, Jr. is with the Department of Electrical Engineering, University of California, Santa Barbara, Calif. 93106 and the Speech Communications Research Laboratory, Santa Barbara, Calif. 93109.

J. D. Markel is with the Speech Communications Research Laboratory, Santa Barbara, Calif. 93109.

## II. A SPECTRAL-FLATNESS MEASURE FOR SPEECH

Let $\{e_n\}$ represent a finite energy real time sequence, so that its $z$ transform, $E = E(z)$, is analytic on the unit circle. If $r_e(0)$ denotes the energy of the time sequence,

$$r_e(0) = \sum_{n=-\infty}^{\infty} e_n^2 = \int_{-\pi}^{\pi} |E[\exp(j\theta)]|^2 \frac{d\theta}{2\pi} \tag{1}$$

then a normalized log spectrum of the time sequence can be defined by

$$V = V(\theta) = \log \{|E[\exp(j\theta)]|^2 / r_e(0)\}. \tag{2}$$

From (1) it can be seen that the energy of the time sequence is the average of its spectrum, so that it is clear that a perfectly flat or constant spectrum will yield a normalized spectrum of zero. Nonzero values for $V$ will represent deviations from a flat or constant spectrum. In this section possible quantitative measures of such deviations will be considered.

One possible measure of deviation from a perfectly flat spectrum is the mean square of the normalized log spectrum, or any constant times that mean square. For example, one could use as such a measure $\eta(E)$ where

$$\eta(E) = \int_{-\pi}^{\pi} \tfrac{1}{2} V^2(\theta) \frac{d\theta}{2\pi}. \tag{3}$$

Such a measure can be zero only in the case of a flat spectrum. This measure weights positive and negative excursions of the normalized log spectrum equally, for the integrand $V^2/2$ is an even function of $V$. As the peaks of speech log spectra (more precisely, the formants) play a more important role than do the valleys in the perception of speech [3], this measure does not seem to be the most appropriate.

In the analysis of speech it would be preferable to use an integrand that is not symmetric, but more heavily weighs the positive excursions of $V$ than the negative excursions. There are an infinite number of such integrands, but we shall here limit our discussion to one which has been utilized in linear prediction. Itakura [4] utilized a statistical approach to linear prediction and minimized a maximum likelihood ratio, which was expressed as an

integral. With the exception of a factor of 2 and a sign, that integral was of the form

$$\mu(E) = \int_{-\pi}^{\pi} \{\exp[V(\theta)] - 1 - V(\theta)\} \frac{d\theta}{2\pi}. \quad (4)$$

The integrand in this case, $e^V - 1 - V$, has the desired properties mentioned above. Fig. 1 shows the two integrands of (3) and (4) as functions of $V$. The solid line represents the integrand of (4) and the dashed line the integrand of (3). It can be noted that with respect to a perfectly flat spectrum, (4) weighs the positive excursions more heavily than (3) and the negative excursions less heavily than (3). A Taylor expansion of the exponential in (4) directly shows that the two integrands become identical for small values of the normalized log spectrum, $V = V(\theta)$.

The integrals of (3) and (4) both have the property that they can equal zero if and only if the spectrum is flat. Based upon our interest in speech signals and the linear prediction approach to speech analysis, (4) appears preferable.

The expression of (4) can be simplified by noting that as $V(\theta)$ represents the normalized log spectrum of the signal, the average of $e^V$ will be unity, giving the result

$$\mu(E) = -\int_{-\pi}^{\pi} V(\theta) \frac{d\theta}{2\pi}. \quad (5)$$

Viewed in this manner, $-\mu(E)$ is simply the average of the normalized log spectrum. Thus, the apparently complex measure of (4) is actually equivalent to a very simple and well known measure. It has been pointed out [5] that $-\mu(E)$ with a factor of 2 represents the zeroth quefrency of the cepstrum and that $\exp[-\mu(E)]$ represents the ratio of the geometric to arithmetic means of the spectrum.

For purposes of normalization, we shall define a spectral-flatness measure $\Xi(E)$, where

$$\Xi(E) = \exp[-\mu(E)] = \exp\left[\int_{-\pi}^{\pi} V(\theta) \frac{d\theta}{2\pi}\right] \quad (6)$$

where $V(\theta)$ is the normalized log spectrum of (2). With this normalization the spectral-flatness measure, $\Xi(E)$, will lie between zero and one, and equal one only for a perfectly flat spectrum. The relationship between the input and output spectral flatness of the inverse filter will now be derived.

### A. *Spectral-Flatness Transformations*

If an input time sequence whose $z$ transform is $X(z)$ is passed through an all-zero inverse filter of the form

$$A_M = A_M(z) = 1 + \sum_{k=1}^{M} a_{Mk} z^{-k}, \quad (7)$$

then the output time sequence will have a $z$ transform $E = E(z)$ of the form

$$E(z) = X(z) A_M(z). \quad (8)$$

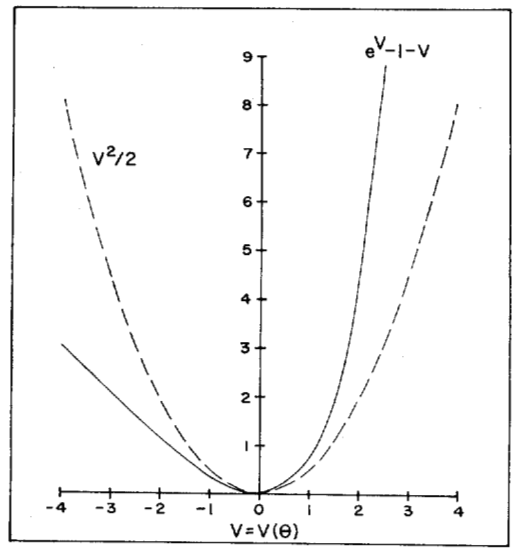


Fig. 1. Weighting of deviations from constant (after Itakura).

If $A_M(z)$ is restricted to having all of its zeros within the unit circle, then $A_M(1/z)$ will be analytic on and within the unit circle and will have all of its zeros outside of the unit circle.

Residue calculus can be applied to show that

$$\int_{-\pi}^{\pi} \log\{|A_M[\exp(j\theta)]|^2\} \frac{d\theta}{2\pi}$$

$$= \int_{-\pi}^{\pi} \log\{|A_M[\exp(-j\theta)]|^2\} \frac{d\theta}{2\pi}$$

$$= 2\,\text{Real}\left(\int_{-\pi}^{\pi} \log\{A_M[\exp(-j\theta)]\} \frac{d\theta}{2\pi}\right)$$

$$= 2\,\text{Real}\left(\oint_{\Gamma} \log[A_M(1/z)] \frac{dz}{2\pi jz}\right)$$

$$= 2\,\text{Real}\{\log[A_M(\infty)]\} = 2\,\text{Real}[\log(1)] = 0, \quad (9)$$

where $\Gamma$ is the unit circle in the $z$-plane. Using this result with (8), one obtains

$$\int_{-\pi}^{\pi} \log\{|E[\exp(j\theta)]|^2\} \frac{d\theta}{2\pi}$$

$$= \int_{-\pi}^{\pi} \log\{|X[\exp(j\theta)]|^2\} \frac{d\theta}{2\pi}.$$

Application of the spectral-flatness measure definition, with $r_e(0)$ and $r_x(0)$ representing the energies of the filter output and input, then leads to the transformation

$$\Xi(E) = \Xi(X) r_x(0)/r_e(0). \quad (10)$$

If the input to the filter $A_M(z)$ is fixed, the only portion of (10) that can produce a change in the output spectral flatness is the term $r_e(0)$, the energy of that output. $\Xi(E)$ will thus be a maximum when $r_e(0)$ is a minimum. Since minimizing $r_e(0)$ is one of the many criteria used to lead

to the autocorrelation method of linear prediction [2], maximizing the spectral-flatness measure of the inverse filter output leads to precisely the same results.

One can utilize standard recursion relations to solve for the inverse filter coefficients in terms of the input autocorrelation sequence. The results are described in terms of what are often called the $k$-parameters, $k_0,k_1,\cdots,k_{M-1}$, whose basic property is that each is less than unity magnitude (see [7]). The energy of the output signal can be found from

$$\alpha_0 = r_x(0) \tag{11a}$$

and

$$\alpha_{m+1} = (1 - k_m^2)\alpha_m, \tag{11b}$$

for $m = 0,1,\cdots,M-1$, which recursively leads to

$$r_e(0) = \alpha_M. \tag{11c}$$

It has been noted [5] that $\exp[-\mu(X)]$ is the ratio of the minimum predictor error energy to the input energy, so that in terms of the coefficients of (6) and (11),

$$\Xi(X) = \lim_{M\to\infty}[r_e(0)/r_x(0)] = \alpha_\infty/\alpha_0. \tag{12}$$

When the optimum inverse filter is used, (10) and (11) lead to the result

$$10\log_{10}\Xi(E) = 10\log_{10}\Xi(X) + 10\log_{10}(\alpha_0/\alpha_M). \tag{13}$$

Thus, not only does the optimum inverse filter maximize the spectral flatness of the output, but it adds to the flatness of the input, on a dB scale, the amount $10\log_{10}(\alpha_0/\alpha_M)$. This amount plays a role in describing a decomposition of the speech log spectrum.

From (8) one can write

$$X(z) = E(z)\frac{1}{A_M(z)}, \tag{14}$$

so that the log spectrum of $X(z)$ is the summation of the log spectra of $E(z)$ and $1/A_M(z)$. It has been shown that the energy of the time sequence associated with $\alpha_M^{1/2}/A_M(z)$ is equal to $r_x(0) = \alpha_0$ [6], so that a direct application of the spectral-flatness measure definition (6) and the result of (9) yields

$$\Xi(1/A_M) = \alpha_M/\alpha_0. \tag{15}$$

Using this with (13) gives the result

$$10\log_{10}\Xi(X) = 10\log_{10}\Xi(E) + 10\log_{10}\Xi(1/A_M). \tag{16}$$

Thus, one can decompose the log spectrum of $X(z)$ in terms of both the log spectra of $E(z)$ and $1/A_M(z)$, and the spectral-flatness measures as indicated in (16).

### *B. Numerical Evaluation*

In order to evaluate a spectral flatness it is necessary to evaluate an integral as in (6), where that integral is the average of the normalized log spectrum. Theoretically one can use the analytic nature of the $z$ transform to show that any zeros of the spectrum must be isolated and the resultant effect will still lead to a finite average for the normalized log spectrum, even though that spectrum may go to minus infinity. For the case of a truncated data sequence a nonzero lower bound can be placed on that integral. It is shown in the Appendix that if the data sequence is limited to $L + 1$ equally spaced points, then its spectral-flatness measure is bounded below by $(L!)^2/(2L)!$. This can be used with (6) to place the lower bound of $\log[(L!)^2/(2L)!]$ on the average of the normalized log spectrum. The integral defining the average is thus theoretically convergent.

A rough estimate of the spectral-flatness measure, in dB, can be found by visually estimating the average dB value from a graph of the normalized log spectrum. A closer estimate can be obtained by using a discrete summation to approximate the integral average. In particular one can use the discrete Fourier transform to obtain discrete samples of the log spectrum which can be normalized by using a numerically calculated energy for the truncated signal. If the frequency samples are spaced closely enough, and there are no zeros in the spectrum, a simple numerical discrete average of the normalized log spectrum can yield very good results. To estimate the accuracy one can increase the frequency resolution. For example, in a typical voiced sound truncated to 128 points, zeros were appended and the spectral-flatness measure was estimated using 256-point, 512-point, and 1024-point fast Fouier tranforms (FFT's). The resulting spectral-flatness measures differed by only 0.1 dB.

If there are any zeros in the spectrum, it becomes necessary to truncate the log spectrum from below so as to avoid a result of minus infinity in the numerical discrete average. More sophisticated numerical quadrature approaches could be utilized for greater accuracy, but we have not found them necessary for speech data. Numerous experiments with actual speech data using different sized discrete Fourier transforms, different truncation limits, and discrete summations to approximate the integrals have demonstrated to us that our present approach to approximating the spectral-flatness measure is reasonable.

For the results to be presented in this paper we have settled upon the use of discrete Fourier transforms having twice as many points as the data sequence. The spectra are truncated from below, so that any value falling below $10^{-5}$ of the peak value is replaced by $10^{-5}$ of its peak, giving a 50 dB dynamic range.

Modifications to this numerical procedure may be necessary for nonspeech sounds, sampling rates outside the range from 6 to 13 kHz, or different filtering before $A/D$ conversion. In particular, if the filtering before $A/D$ conversion has a cutoff frequency significantly below the half sampling frequency, the sampled signal will have a very low spectrum for a portion of the range. The effect of this will be to introduce large negative excursions of the normalized log spectrum. This will result in a decreased spectral-flatness measure, which we shall show is undesirable, and may also result in numerical difficulties in

estimating that measure. These difficulties can be taken care of by increasing the number of points used in the discrete Fourier transform and also increasing the allowable dynamic range by lowering the truncation value. For reasons which will become obvious, we consistently prefilter the analog signal with a sharp cutoff at (not below) the Nyquist frequency, and do not eliminate any low frequencies that can pass through a high quality microphone.

In the experimental results to be presented, the approach just described is used to approximate the spectral-flatness measures of the input, $\Xi(X)$, to the inverse filter. Equation (15) is used to give the spectral-flatness measures for the reciprocal inverse filters, $\Xi(1/A_M)$, and (16) to give the spectral-flatness measures of the inverse filter outputs $\Xi(E)$ in terms of $\Xi(X)$ and $\alpha_M/\alpha_0$.

If the data sequence is differenced before the spectral-flatness measure approximation, the zero introduced at zero frequency may lead to numerical difficulties. Two approaches can be utilized in such a case. The least accurate would be to simply ignore the zero frequency term in computing the numerical discrete average of the normalized log spectrum. It is preferable to evaluate the approximation for the spectral-flatness measure before differencing, and then to apply (10) after filtering with the filter $A_1(z) = 1 - z^{-1}$.

### C. Experimental Results

To illustrate the flattening effect of the inverse filter, results from an unvoiced and a voiced sound are shown in Figs. 2 and 3. These isolated analysis frames were obtained from the utterance discussed in Section III of this paper. Each was sampled at a 6.5 kHz rate, and a 128-point time window was applied. A Hamming window was introduced before analysis. Each part of the figures is labeled with two numbers, the filter order $M$ and a spectral-flatness measure in dB, defined by

$$SF(\cdot) = 10 \log_{10} \Xi(\cdot).$$

The parts of the figures each show normalized log spectra, with the inverse filter output on the left and the reciprocal inverse filter on the right represented by $E(z)$ and $1/A_M(z)$ respectively. The grid marks on the left of each figure indicate 5 dB differentials.

For $M = 0$, the normalized log spectrum of $E(z)$ equals that of the original signal and that of $1/A_M(z)$ equals zero. As $M$ increases, the flatness of $E(z)$ increases, both qualitatively and quantitatively. For the largest $M$ used, $M = 300$, it can be noted that almost all of the spectral-flatness measure has been transferred from $E(z)$ to $1/A_M(z)$ and $1/A_M(z)$ has a spectrum that is essentially identical to the original signal. In both figures one can note that $1/A_8(z)$ has a log spectrum that represents a visually smoothed version of the original with clearly visible formant structure. These figures illustrate the spectral flattening properties of the inverse filter and show the representative numerical spectral-flatness measures for the two important classes of unvoiced and voiced sounds.

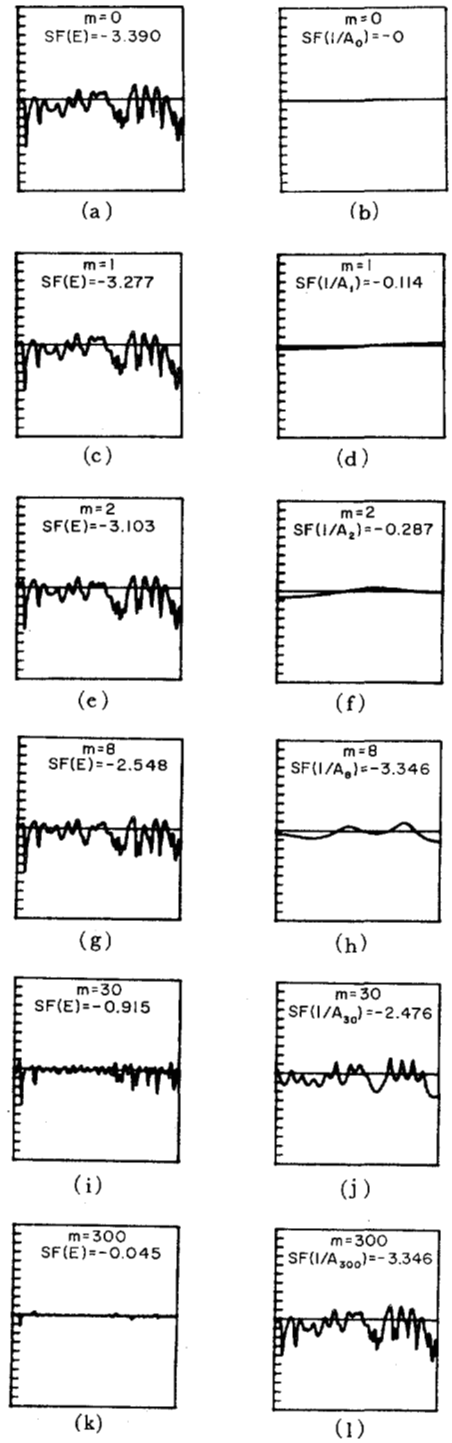


Fig. 2. Unvoiced flattening example.

## III. SPECTRAL-FLATNESS OF TWO DRIVING FUNCTION MODELS

In the autocorrelation method of linear prediction, the inverse filter output is considered to be representative of the driving function to the vocal tract with glottal and lip radiation shaping removed. The driving function is modeled as random noise for unvoiced sounds and equally spaced unit samples for voiced sounds. The inverse filter $A_M(z)$ is chosen so as to minimize the energy output, or equivalently, maximize the output spectral flatness. One point of interest is the desired theoretical maximum of that

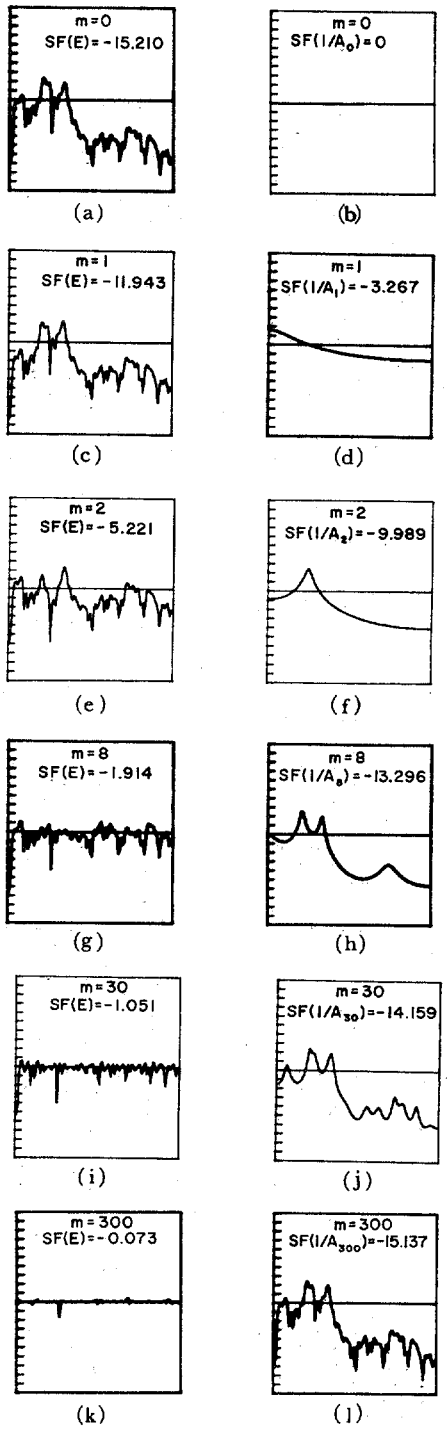


Fig. 3. Voiced flattening example.

output spectral flatness. Increasing the order of the filter, $M$, can increase the spectral flatness beyond a desired theoretical point.

The order of the filter has been chosen by numerous criteria. Markel [2] has empirically found that for reasonable modeling of the spectral resonance behavior, using the log spectrum of $1/A(z)$, there should be roughly one complex pole pair to span each 700 Hz range out to the Nyquist frequency. Thus, for a case where the Nyquist frequency is 3.25 kHz, one needs 4 or 5 complex pole pairs, or an $M$ of 8–10. Wakita [7] and Atal and Hanauer [1] have related the order of the inverse filter to the dimensions of the vocal tract, and have illustrated that if $\Delta t$ is the sample time, $\Delta t = 1/F_s$, then $M\Delta t$ should represent the time that it takes an acoustical signal to transverse the vocal tract twice. This then can be used to relate sampling frequency, vocal tract length, velocity of sound, and filter order. In each of these viewpoints the choice of $M$ is proportional to the sampling frequency, and in order to compare results for identical analog speech signals sampled at different rates one should keep the ratio $M/F_s$, where $F_s$ is the sampling frequency, constant.

The desired theoretical spectral-flatness measures at the output of the inverse filter will now be derived for unvoiced and voiced sounds.

### A. Ideal Unvoiced Driving Function Model

One possible model in the case of unvoiced fricative sounds for the driving function is uncorrelated Gaussian noise. Gray [8] has shown that the log spectrum of such a signal will have an expected value that is less than the logarithm of the expected value of the spectrum by an amount $\gamma$, where $\gamma$ is Euler's constant $0.5772\cdots$, at least at the sample frequency points other than zero and the Nyquist frequency. Thus a numerically evaluated flatness measure will have an expected value of roughly $\exp(-\gamma)$ or $-2.5$ dB. It should be pointed out that this result does not depend upon the sampling frequency or the length of the time window used for analysis.

### B. Ideal Voiced Driving Function Model

A driving function for voiced sounds is the set of $L + 1$ equally spaced samples

$$y_k = 0 \text{ for } k \neq l, l + P, l + 2P, \cdots, l + LP. \quad (17)$$

Here $y_l$ is the first sample in the time window, $y_{l+LP}$ is the last, and $P$ represents a pitch period. It is shown in the Appendix that the spectral flatness will lie between $(L!)^2/(2L)!$ and one, equalling the former if the time sequence takes on values distributed according to the binomial coefficients as in (A11) or (A12), and the latter if all but one of the samples equals zero. In addition, it is shown that if all the samples are of the same size, then the spectral-flatness measure equals $1/(L + 1)$.

As a result, if there is only one such sample in the time window, the resulting spectral-flatness measure is 1, or 0 dB. If there are two such samples ($L = 1$) the measure will lie between 1/2 and 1, or $-3$ dB and 0 dB, being closer to $-3$ dB if the samples are of almost the same size. If there are three such samples the measure will lie between 1/6 and 1, or $-7.8$ dB and 0 dB, with the value 1/4 or $-6$ dB when the samples are of roughly the same size.

### C. Experimental Results

The example utterance "Will the rest follow soon?" from the first speaker on the 1972 International Speech Communication and Processing Conference tape is used to illustrate the preceding material. The sampling frequency $F_s$ was 6.5 kHz in all but one of the examples used (see Table I), and in each case the analog data was sharply

TABLE I

| Example | Sampling Frequency $F_s$ | Number of Samples $N$ | Window Length | Type of Window |
|---|---|---|---|---|
| Fig. 5(a) | 6.5 kHz | 128 | 19.69 ms | rectangular |
| Fig. 5(b) | 6.5 kHz | 128 | 19.69 ms | Hamming |
| Fig. 5(c) | 13.0 kHz | 256 | 19.69 ms | Hamming |
| Fig. 5(d) | 6.5 kHz | 256 | 39.38 ms | Hamming |
| Fig. 6(a) | 6.5 kHz | 128 | 19.69 ms | rectangular |
| Fig. 6(b) | 6.5 kHz | 128 | 19.69 ms | Hamming |

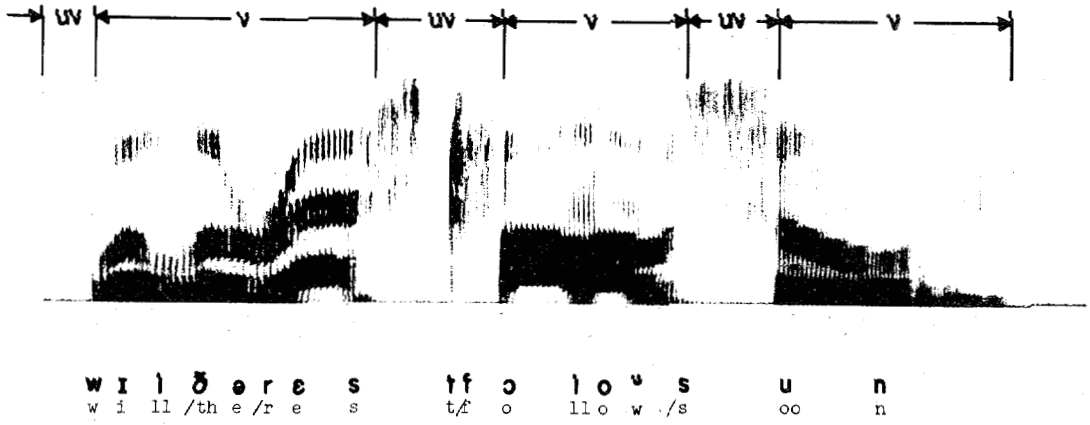


Fig. 4. Spectrogram of utterance "Will the rest follow soon."

prefiltered at the Nyquist frequency, $F_s/2$. A spectrogram of the frequency range 0 through 3.25 kHz, made by playing the digital samples through a digital-to-analog converter is shown in Fig. 4. The voiced portions are indicated by $V$ and the unvoiced portions are indicated by $UV$. In addition, the approximate time location of the speech sounds are indicated along with phonemic symbols.

Table I summarizes the differences between the various analysis parameters, only Fig. 5 pertains to this section. Each data window is made up of $N$ points resulting in a time window of length $N/F_s = N\Delta t$. The time between windows is fixed in all examples at 29.538 ms, so only in the case of Fig. 5(d) will adjacent windows overlap.

In each part of the figures two curves are shown, the lower representing the spectral-flatness measures at the input to the inverse filter, $10 \log_{10} \Xi(X)$, and the upper representing the spectral-flatness measures at the outputs, $10 \log_{10} \Xi(E)$. Calculations were effected as described in Section II. The value of $M$ used is 8 in the cases using a sampling rate of 6.5 kHz and 16 when the sampling rate is 13 kHz.

A number of important conclusions can be drawn from these four examples. The spectral-flatness measure of the inverse filter output varies far less than that of the input. During unvoiced portions the theoretical model predicted an average level of $-2.5$ dB which compares quite well with the experimental results indicated in the figure even with varying sampling frequency and window length.

During voiced portions, the theoretical model predicted a wide range of values, depending upon the number of samples in the model (number of pitch periods in a time window) and the relative sizes of these samples. For the case of two such samples, the predicted spectral-flatness measure ranged from $-3$ dB to 0 dB, taking on the value of $-3$ dB when the samples were of equal size. With the exception of the end of the third voiced intervals in Fig. 5(a)–(c), the average number of pitch periods per analysis window was about two, and the experimental flatness measures averaged about $-3$ dB. This compares well with the theoretical predictions, particularly in the light of the overly simplified model, for the actual inverse filter output is certainly more complex than a pair of equal samples.

During the end of the third voiced interval, the pitch period decreases, resulting in a decrease in the spectral flatness of the inverse filter outputs. This result is expected from the theoretical model, but the actual amount of decrease is difficult to predict in advance, for in each frame it depends upon the number of pitch samples in the model for that time window, and their relative size. In Fig. 5(d) the number of pitch periods per analysis window is twice that of the earlier examples, and as expected for voiced portions of speech, the spectral-flatness measure is decreased.

Looking to the spectral flatness of the inverse filter inputs, the lower curves, one can note a dramatic difference between the measures for voiced and unvoiced sounds. In this case essentially perfect voiced/unvoiced decisions might be made by using this difference, but this approach will often be incorrect in the case of unvoiced plosive sounds.

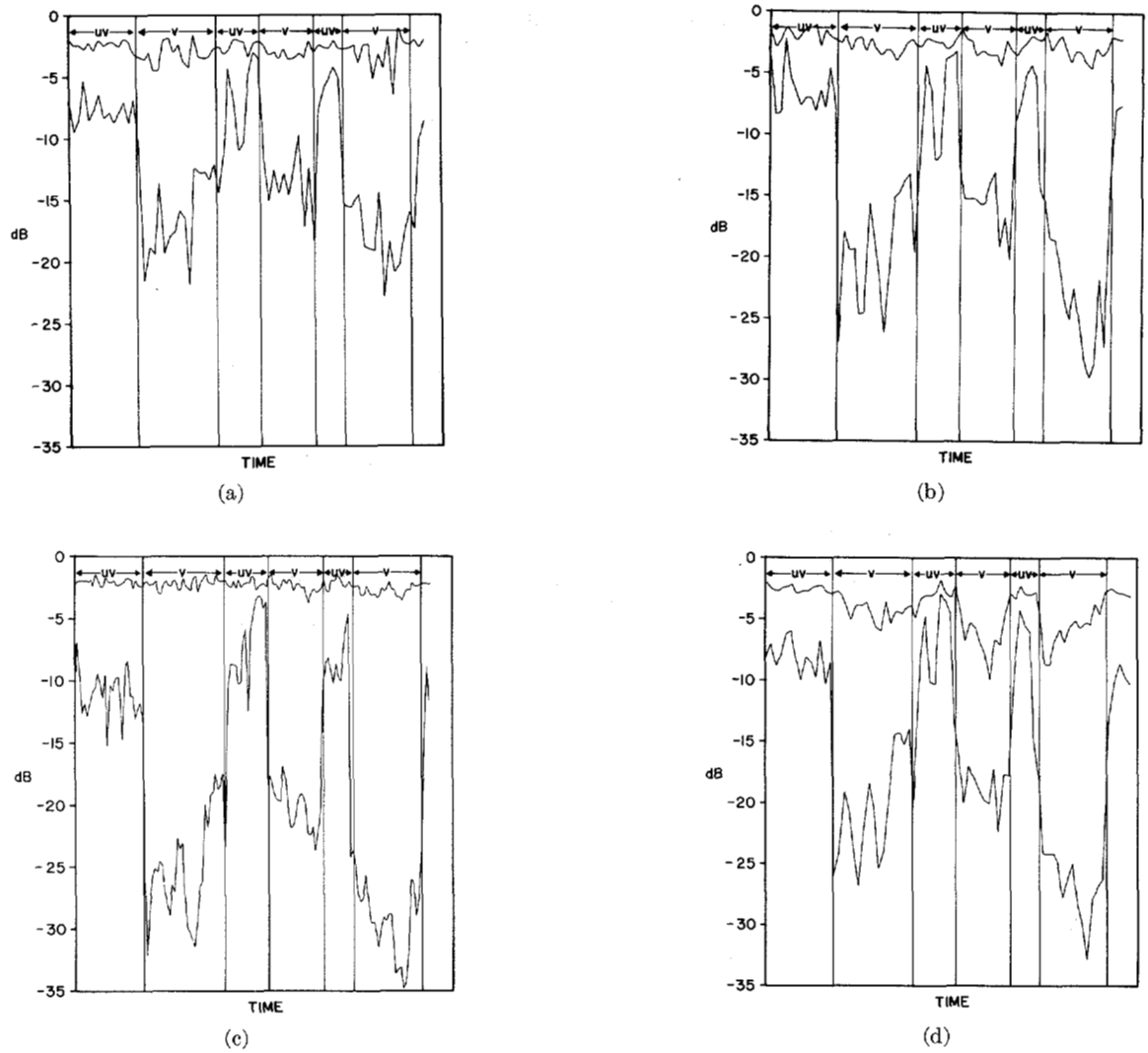


Fig. 5. Spectral flatness results: (a) rectangular window, (b) Hamming window, (c) increased sample frequency, and (d) increased window length.

A comparison of Fig. 5(a) and (b) shows that the spectral-flatness measure of the windowed speech signal is decreased during voiced portions by the use of a Hamming window. The window is utilized because it has been experimentally and theoretically noted [5] that for the autocorrelation method, improved spectral representation is obtained.

A comparison of Fig. 5(b) and (c) shows that increasing the sampling rate (keeping the same time window) has the effect of reducing the spectral-flatness measure of voiced sounds.

These results give further justification for the spectral-flatness measure usage, in that results obtained are consistent with what one knowledgeable in spectral characteristics of speech would expect. For example, a nasal sound such as one at the end of the utterance will generally have a larger negative spectral slope and thus a smaller spectral-flatness measure than other voiced sounds. This fact is clearly seen in Fig. 5(b) where the smallest spectral flatness over the utterance is in the nasal region.

## IV. SPECTRAL FLATNESS AND ILL-CONDITIONING

In solving for the coefficients of the inverse filter, one must solve a set of $M$ simultaneous algebraic equations. The matrix of coefficients of these equations is the $M$ by $M$ autocorrelation matrix $\boldsymbol{R}$ whose elements $r_{ik}$ for $i = 1,2,\cdots,M$ and $k = 1,2,\cdots,M$ are given by the autocorrelation values of the inverse filter input sequence, $r_{ik} = r_x(i-k)$. One might consider this as an implicit inversion of the matrix $\boldsymbol{R}$, yet in actual fact one does not need the matrix inverse itself, only the resultant vector $\boldsymbol{R}^{-1}\boldsymbol{r}$, where $\boldsymbol{r}$ is the column matrix or vector whose elements are $r_x(1),r_x(2),\cdots,r_x(M)$.

There are numerous "measures" of ill-conditioning of matrices. The most common of these are the $N$ and $M$ condition numbers of Turing and the $P$ and $H$ condition numbers of Todd [9]–[11]. The last two are identical for the case of symmetric matrices, and equal the ratio of the

maximum and minimum eigen values of the matrix (for the case of real positive eigen values). This ratio has sometimes been called simply "the" condition number. Ekstrom [12] has used it and a theorem of Szegö's [13] to show that for large $M$ this condition number approaches the dynamic range of the spectrum, the ratio of the maximum to minimum values thus illustrating why increasing of the sampling rate can add to ill-conditioning. This approach must of necessity exclude possible spectral zeros which yield infinite dynamic range.

In analysis of voiced speech there are often fairly deep negative excursions of the normalized log spectrum due to the periodicity of the speech samples. Experimentally, however, in our analysis of thousands of frames of actual speech data, these alone have not led to ill-conditioning when the autocorrelation equations are solved recursively using Robinson's or Levison's method. We have therefore been motivated to introduce a more elementary measure of ill-conditioning, which more closely corresponds with experimental results. This measure will be defined as a number which lies between zero and one, taking on the value of one for a perfectly conditioned problem, when $\boldsymbol{R}$ is the identity matrix, and zero for a singular problem, when $\boldsymbol{R}$ is a singular matrix.

One elementary approach is to utilize a normalized determinant of the matrix in question. It has been shown [4] that if the determinant of $\boldsymbol{R}$ is divided by the $M$th power of its largest element, $r_x(0)$, then

$$|\boldsymbol{R}|/r_x^M(0) = |\boldsymbol{R}|/\alpha_0^M = \prod_{m=0}^{M-1} (\alpha_m/\alpha_0) \qquad (18)$$

where the $\alpha_m$ are found recursively from (11). As the ratio $(\alpha_m/\alpha_0)$ forms a decreasing sequence with increasing $m$, going from one at $m = 0$ to $\Xi(X)$ as $m$ goes to infinity, the ratio of (18) can become very small as $M$ increases. We have experimentally observed that in solving for $\boldsymbol{R}^{-1}\boldsymbol{r}$, ill-conditioning is only slightly worse for a very large matrix than for one of lower order. For example, in Fig. 3(1) no difficulty was observed in solving the equations with single precision arithmetic even though measure of singularity in (18) is on the order of $10^{-300}$. The ill-conditioning in such cases appears to taper off to a constant, and as a result we shall modify (18) by using its $M$th root as a measure,

$$\rho_M = |\boldsymbol{R}|^{1/M}/\alpha_0 = \left[\prod_{m=0}^{M-1} (\alpha_m/\alpha_0)\right]^{1/M}. \qquad (19)$$

This expression for $\rho_M$ indicates that $\rho_M$ represents the geometric mean of the decreasing sequence $(\alpha_m/\alpha_0)$. It will decrease from 1, for $M = 1$, and approach the limiting value

$$\lim_{M\to\infty} \rho_M = \rho_\infty = \alpha_\infty/\alpha_0 = \Xi(X). \qquad (20)$$

The spectral-flatness measure is thus both a lower bound and a limiting value of $\rho_M$, and as such can itself be considered a measure of ill-conditioning. This result leads to a more quantitative explanation for a number of well known "rules of thumb" as follows:

1) It takes more accuracy to analyze voiced sounds than unvoiced.

2) The use of a Hamming or Hanning window increases the amount of computational accuracy needed.

3) Increasing the sampling rate increases the amount of computational accuracy needed.

4) Proper preemphasis or prewhitening can decrease the amount of computational accuracy needed.

Using the spectral-flatness measure as an indicator of ill-conditioning, the first three of these comments have already been illustrated in Fig. 5. In a qualitative manner one can note that the voiced sounds are less flat than the unvoiced sounds, the window decreases the flatness of the voiced sounds, and the increased sampling rate decreases the flatness of the voiced sounds.

It is seen that the normalized log spectrum of the voiced sound in Fig. 3 has a decreasing trend for increasing frequency. If the Nyquist frequency were increased by higher sampling of the actual acoustic waveform the trend would continue downward, resulting in a less flat spectrum and lower spectral-flatness measure. The spectral-flatness measure can be improved upon by some preemphasis of the signal, before analysis by the inverse filter. Such preemphasis can result in reducing ill-conditioning, as discussed in the following section.

## V. PREEMPHASIS OF THE SPEECH DATA

Markel [2] has noted that differencing the input speech signal is a simple and effective way of accenting the higher formants, thus allowing more accurate formant tracking results. Wakita [7] has observed that differencing (or other such preemphasis which results in an approximate 6 dB/octave filter slope) is necessary to obtain reasonable vocal tract area function shapes when using the inverse filter model on voiced sounds. This result is due to the linear speech model, which has the vocal tract spectral shape modified by approximately $-12$ dB/octave from the glottal wave shape and $+6$ dB/octave from the lip radiation characteristics. The net result, $-6$ dB/octave, is approximately cancelled by differencing the input samples before analysis.

We have recently observed, experimentally, that the probability of numerically caused instabilities in the filter $1/A_M(z)$ is greatly reduced by preemphasis or prewhitening of the speech data, and a three to four bit reduction in the needed accuracy when the calculations are carried out finite word length fractional arithmetic [14]. Thus we conclude that the rather trivial modification of preemphasis is of fundamental importance to inverse filter analysis techniques, and we speculate that this will also be true for other digital analysis techniques.

From the results of the preceding section, one can note that a low spectral-flatness measure can indicate possible ill-conditioning. This suggests the possibility of preemphasis of the speech signal before inverse filtering.

One approach to preemphasis is to utilize a low-order inverse filter, and maximize the spectral flatness of its output. The preemphasis filter is thus found by the same criteria as is the inverse filter. We have had considerable success using a simple first order preemphasis filter, and will discuss its properties in greater detail.

### A. First-Order Preemphasis

A first-order preemphasis filter would be of the form $1 - \mu z^{-1}$. A differencer would use $\mu = 1$. We shall show that the optimal preemphasis filter which maximizes the output spectral-flatness measure will have $\mu = r_s(1)/r_s(0)$, where $\{r_s(n)\}$ represents the autocorrelation sequence for the input speech data sequence $\{s_n\}$. The sequence $\{x_n\}$ will still represent the input to the inverse filter, $A_M(z)$, and will equal the output of the preemphasis filter when no windowing takes place between the two. If $\{f_n\}$ is the time sequence of the preemphasis filter output then (10) can be applied to give

$$\Xi(F) = \Xi(S) r_s(0)/r_f(0), \tag{21}$$

and a direct evaluation gives $r_f(0)$ as

$$r_f(0) = (1 + \mu^2) r_s(0) - 2\mu r_s(1). \tag{22}$$

$\Xi(F)$ will be a maximum when $r_f(0)$ is a minimum, which occurs for $\mu = r_s(1)/r_s(0)$, and that maximum is $\Xi(S)/(1 - \mu^2)$. Experience has shown us that low spectral values for speech data usually arise in the case of voiced sounds where $r_s(1)$ is apt to be very near $r_s(0)$. In such cases this preemphasis can greatly increase the spectral flatness measure and the preemphasis filter becomes almost a differencer. For most unvoiced sounds $r_s(1)/r_s(0)$ is relatively small and the preemphasis filter has little effect.

Even when the preemphasis filter is only approximately optimal, the spectral-flatness measure can be enhanced. In particular, one can use (22) to show that any value of $\mu$ that lies between zero and twice the optimal value will improve the spectral-flatness measure. We have experimentally found that 1-bit quantization of $\mu$ is superior to constant preemphasis and only slightly inferior to direct application of $\mu$.

### B. Experimental Results

Fig. 6 illustrates the results of the application of optimal preemphasis windows. Table I gives the analysis conditions under which the graphs were obtained. The time window follows the preemphasis filters and precedes the inverse filter. Spectral-flatness measures at the input and output of the inverse filter were evaluated as described in Section II. To keep the total order of the filters the same, the order of the inverse filter, $M$, was reduced by the order of the preemphasis filter. This allows a more direct comparison with the results of Fig. 5.

Fig. 6(a) illustrates the application of first-order preemphasis with no Hamming window. By a direct comparison with Fig. 5(a), one can see that the input spectral flatness is greater while the output spectral flatness is slightly less. This is to be expected, for the preemphasis filter increases the spectral flatness at the input of the inverse filter. The combination preemphasis filter and inverse filter is eighth order in Fig. 6(a), and thus cannot have an output spectral flatness which is as good as that of the optimum eighth-order inverse filter shown in Fig. 5(a). Thus one might think of the total system, preemphasis and inverse filter, as a suboptimal eighth-order filter.

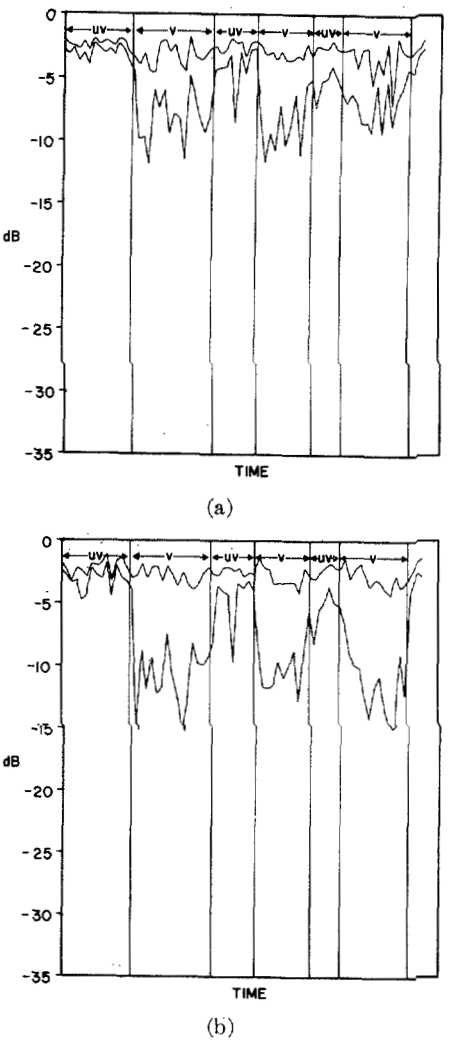


Fig. 6. Spectral flatness with preemphasis: (a) rectangular window and (b) Hamming window.

In addition, the distance between the lower and upper curve, $10 \log_{10} \alpha_M/\alpha_0$ [from (13)] indicates that the ill-conditioning of the solution process is considerably reduced [from (18) and (19)].

The results of the application of the window between preemphasis and inverse filtering are shown in Fig. 6(b). A comparison of Fig. 6(a) and (b) shows that the spectral flatness at the output of the inverse filters is essentially unchanged in its overall behavior; however, as in the case of Fig. 5, the spectral flatness during voiced portions is decreased, with the largest change once again for the nasal /n/ region.

## VI. SUMMARY

A spectral-flatness measure has been developed to help give insight into the whitening process of the autocorrelation method of linear prediction. It assigns a numerical value from 0 to 1, or $-\infty$ dB to 0 dB, to each speech spectrum. Only a perfectly flat or constant spectrum can have a flatness of 0 dB.

Spectral-flatness measures for theoretical model driving functions were compared with those of actual inverse filter outputs resulting from real speech data. It was shown that the spectral-flatness measure is closely correlated with possible ill-conditioning of the analysis problem—the lower the spectral flatness the more ill-conditioned the problem. Preemphasis of the speech signal by means of a one-term linear predictor was shown to greatly enhance the spectral flatness of the signal and make the ill-conditioning less significant.

Recent research [14] with fixed-point solution of the inverse filter has further demonstrated the applicability of the spectral-flatness measure for predicting ill-conditioning as a function of the type of sound, sampling frequency, and preemphasis conditions.

## APPENDIX

## LOWER BOUNDS ON SPECTRAL-FLATNESS MEASURES

If a data sequence is limited in the number of points, a nonzero lower bound on spectral flatness can be obtained. Let $Y(z)$ be the $z$ transform of a data set which consists of $L+1$ samples equally spaced by $P$ so that

$$Y(z) = \sum_{\nu=0}^{L} y_{l+\nu P} z^{-(l+\nu P)}, \tag{A1}$$

where $l$ is any integer. It is a simple matter of applying the definition of spectral flatness to show that $\Xi(Y)$ is independent of $P$, for integer nonzero $P$, so that without loss of generality we take $P$ to be one.

As the spectral-flatness measure depends only upon the normalized log spectrum, there will always be a polynomial whose roots are on or within the unit circle of the form

$$G(z) = \sum_{\nu=0}^{L} g_\nu z^{-\nu} \text{ with } g_0 = 1 \tag{A2}$$

such that

$$\Xi(Y) = \Xi(G). \tag{A3}$$

As $G(z)$ has all of its zeros on or in the unit circle, one can show that

$$|g_\nu| \leq \frac{L!}{\nu!(L-\nu)!} \tag{A4}$$

with equality holding for any $\nu > 1$ if and only if

$$G(z) = [1 \pm z^{-1}]^L. \tag{A5}$$

By using the fact that $G(z)$ has all of its zeros in or on the unit circle and that $G(z)$ is described by (A2), one can show that

$$\Xi(G) = [r_g(0)]^{-1} \tag{A6}$$

with

$$r_g(0) = \sum_{\nu=0}^{L} g_\nu^2. \tag{A7}$$

From (A7) and the bound of (A4), one finds

$$r_g(0) \leq \frac{(2L)!}{(L!)^2} \tag{A8}$$

with equality holding if and only if $G(z)$ is given by (A5). This can then be applied to (A6) to give a lower bound of $\Xi(G)$, which can be obtained only when all of the zeros of $G(z)$ are equal to either plus one or minus one.

Working backwards to $Y(z)$ of (A1), this leads to the conclusion that if $Y(z)$ is given as shown, then

$$\Xi(Y) \geq \frac{(L!)^2}{(2L)!} \tag{A9}$$

with equality holding if and only if $Y(z)$ is given as

$$Y(z) = y_l[1 \pm z^{-P}]^L \tag{A10}$$

so that $y_{l+\nu P}$ can be given by either

$$y_{l+\nu P} = y_l \frac{L!}{\nu!(L-\nu)!} \qquad \nu = 0,1,2,\cdots,L \tag{A11}$$

or

$$y_{l+\nu P} = y_l \frac{L!}{\nu!(L-\nu)!}(-1)^\nu \qquad \nu = 0,1,2,\cdots,L. \tag{A12}$$

The spectral-flatness measure of equal sized samples can be found from (A6) and (A7) by letting each of the $g_\nu$ approach unity. This leads to a spectral-flatness measure of $1/(L+1)$.

## REFERENCES

[1] 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, Aug. 1971.
[2] J. D. Markel, "Digital inverse filtering—a new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.
[3] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. New York: Springer, 2nd ed., 1972.
[4] F. Itakura, "Speech analysis and synthesis systems based on statistical method" (in Japanese), Ph.D. Dissertation, Department of Engineering, Nagoya Univ., Nagoya, Japan, Mar. 1972.
[5] J. I. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt, Beranek and Newman, Inc., Cambridge, Mass., Rep. 2304.
[6] J. D. Markel and A. H. Gray, Jr., "On autocorrelation equa-

tions as applied to speech analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 69–79, Apr. 1973.
[7] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, Oct. 1973.
[8] A. H. Gray, Jr., "Log spectra of Gaussian signals," *J. Acoust. Soc. Amer.*, to be published.
[9] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*. San Francisco: Freeman, 1963.
[10] A. M. Turing, "Rounding-off errors in matrix processes," *Quart. J. Mech. and Appl. Math.*, vol. 1, pp. 287–308, 1948.
[11] J. Todd, "The condition of certain matrices. I," *Quart. J. Mech. Appl. Math.*, vol. 2, pp. 469–472, 1949.
[12] M. P. Ekstrom, "A spectral characterization of the ill-conditioning in numerical deconvolution," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 344–348, Aug. 1973.
[13] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications*. Berkeley, Calif.: Univ. of Calif. Press, 1958.
[14] J. D. Markel and A. H. Gray, Jr., "Fixed-point truncation arithmetic implementation of a linear prediction autocorrelation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.

# A Programming System for Studies in Speech Synthesis

P. V. S. RAO AND R. B. THOSAR

*Abstract*—**This paper describes a speech synthesis system which is particularly suitable for experimental investigations. The synthesis is accomplished in two stages. The concatenation stage generates a schematized spectrographic representation corresponding to the symbolic input. The second stage consists in generating the corresponding acoustic signal. The steady state characterization of each phoneme is supplied as data. Independent concatenation procedures incorporate context dependent effects such as format transitions, changes in the normal duration of vowels, etc. The parameter values for these procedures are obtained by a set of rules. Applicability of a rule is determined by attributes assigned to the phonemes.**

**The phonemes are divided into classes and subclasses by the attribute assignment. The attribute STOP, for instance, defines the class of all stop consonants and BILABIAL STOP would define the set /p, b, m/. Thus, a rule specifies a parameter value when a subclass of phonemes occur, in the context of another subclass. Such a formulation considerably reduces the number of rules. The classification as well as the rules are supplied as data to the system, giving it considerable flexibility.**

**The spectrographic output of the concatenation stage is used to actuate a simulated series terminal analog synthesizer. Rudimentary prosodics are incorporated which modify a monotonous pitch contour with stress markers and interrogative or declarative termination of a sentence.**

## INTRODUCTION

HIGH quality synthetic speech, almost indistinguishable from natural, can be produced using low data rate terminal analogue synthesisers [12], [13]. This fact warrants a closer look at synthesis-by-rule which attempts to generate the formant contours from a symbolic representation of the message to by synthesized. The central problem in synthesis-by-rule is the generation of contextual effects dependent primarily on immediate adjacent phonemes. The so-called coarticulation effects [1] may also contribute to the naturalness of synthetic speech, though they do not seem to be significant for intelligibility [2].

Most speech synthesis systems concentrate on faithful reproduction of vowel formant transitions in consonantal context. The transitions are generated by assuming a suitable model. Thus, Holmes *et al.* [3] specify the formant loci and compute the transitions by linear interpolation. Rabiner [5] uses the solution of a second degree differential equation to compute the formant contours from specified time constants. Clearly, a large number of parameters require to be specified to generate all the vowel–consonant pairs. Holmes *et al.* [3] reduce this requirement by rank ordering the set of sounds, but this necessitates rather arbitrary enlargement of the basic set of sounds to account for more complex interactions [4].

Besides formant transitions, other systematic contextual effects have also been observed, which must be considered in synthesis-by-rule. Vowel durations vary according to the contiguous consonent (House, 1958). The variations in steady state formants of liquids and glides are important for their perception [7], [9]. The fricatives show evidence of transitions in consonant clusters [10].

In this paper we propose a programming system for speech synthesis which would allow incorporation of these effects in a systematic manner. Concatenation rules, the basic set of sounds and context effects can be modified or added to the system in a relatively painless way. The system would, thus, serve as a flexible tool for studies in speech synthesis.

In the following section we outline the synthesis strategy in a general way. Section B describes some details of

Manuscript received March 28, 1973; revised August 28, 1973 and December 12, 1973.
The authors are with the Tata Institute of Fundamental Research, Bombay, India.