Which is the Better Entropy Expression for Speech Processing: $-S \log S$ or $\log S$?

RODNEY W. JOHNSON AND JOHN E. SHORE, SENIOR MEMBER, IEEE

Abstract—In maximum entropy spectral analysis (MESA), one maximizes the integral of $\log S(f)$, where $S(f)$ is a power spectrum. The resulting spectral estimate, which is equivalent to that obtained by linear prediction and other methods, is popular in speech processing applications. An alternative expression, $-S(f)\log S(f)$, is used in optical processing and elsewhere. This paper considers whether the alternative expression leads to spectral estimates useful in speech processing. We investigate the question both theoretically and empirically. The theoretical investigation is based on generalizations of the two estimates—the generalizations take into account prior estimates of the unknown power spectrum. It is shown that both estimates result from applying a generalized version of the principle of maximum entropy, but they differ concerning the quantities that are treated as random variables. The empirical investigation is based on speech synthesized using the different spectral estimates. Although both estimates lead to intelligible speech, speech based on the MESA estimate is qualitatively superior.

I. Introduction

MANY spectral analysis techniques start with measured values of the autocorrelation function $R(t)$ of a signal at a set of points. One class of techniques proceeds by extrapolating $R(t)$ to reasonable values at the unknown points. The extrapolated autocorrelation function is equivalent to a power spectrum estimate, since the power spectrum $S(f)$ of a band-limited stationary process is related to its autocorrelation function by a Fourier transform.

Perhaps the best known extrapolation technique is Burg's maximum entropy spectral analysis (MESA) [1], [2], in which the power spectrum $S(f)$ is estimated by maximizing

$$\int_0^W \log S(f)df \tag{1}$$

subject to the constraints

$$R_r = R(t_r) = \int_{-W}^{+W} S(f)\exp(2\pi i t_r f)df \tag{2}$$

where W is the bandwidth and where $R(t_r)$, $r = 1, 2, \cdots, M$, are known values of the autocorrelation function. The MESA estimate of $S(f)$ has the well-known all-pole, autoregressive, or linear prediction form, which can also be derived by various equivalent formulations [3]-[6]. It has become one of the most widely used spectral analysis techniques in geophysical data processing [7]-[9] and speech processing [4], [10].

"Maximum entropy spectral analysis" is also used in image processing. In that field, however, the phrase refers not only to successful estimates produced by maximizing (1) [11]-[13], but also to estimates produced by maximizing [14]-[16]

$$-\int_0^W S(f)\log S(f)df. \tag{3}$$

Spectral estimates based on (3) have also been studied for ARMA and meteorological time series [17], [18]. Although there is controversy in the image processing literature about whether (1) or (3) yields better estimates [16], [19], the success of (3) in image processing raises the question of whether (3) might also be useful in speech processing. We consider the question in this paper, and we attempt to answer it. As part of our investigation, we also derive a generalization of the estimate produced by maximizing (3), one that takes into account a prior estimate of the unknown power spectrum.

Our paper is organized as follows. In Section II we review derivations of the forms (1) and (3), and we discuss theoretical arguments for each of them. We then turn to an empirical comparison. Our approach is discussed in Section III and the results are summarized in Section IV. A general discussion then follows in Section V.

II. Background

In this section we give brief derivations of the spectral estimators that result from maximizing (1) and (3). We show that both estimators result from applying the principle of minimum cross-entropy [20]-[24], a generalized form of the principle of maximum entropy [25]-[27]. However, they differ concerning the quantities that are treated as random variables. In the case of (1), the underlying random variables are the coefficients of a Fourier series model and the spectral powers $S(f)$ are expected values. In the case of (3), the spectral power $S(f)$—suitably normalized—is treated as a probability density and the underlying random variable is the frequency.

Cross-entropy minimization estimates an unknown probability density $q^\dagger(x)$ from a prior estimate $p(x)$ and known expected values

$$\int q^\dagger(x)g_r(x)dx = \overline{g}_r \tag{4}$$

Manuscript received February 24, 1983; revised July 29, 1983.

The authors are with the Computer Science and Systems Branch, Information Technology Division, Naval Research Laboratory, Washington, DC 20375.

$(r = 0, \cdots, M)$. The estimate is obtained by minimizing the cross-entropy

$$H(q, p) = \int q(x) \log \left[\frac{q(x)}{p(x)} \right] dx \tag{5}$$

subject to the constraints (4) and

$$\int q(x) dx = 1. \tag{6}$$

The resulting estimate of $q^\dagger(x)$ has the form [20], [22], [28]

$$q(x) = p(x) \exp \left[-\lambda - \sum_{r=0}^{M} \beta_r f_r(x) \right] \tag{7}$$

where the β_r and λ are Lagrangian multipliers determined by (4) and (6). Cross-entropy minimization reduces to entropy maximization when the prior $p(x)$ is uniform.

A. The $-\log S$ Form

In deriving MESA, Burg's approach was to extrapolate $R(t)$ in a manner that maximizes the entropy of the underlying stochastic process [1], [2]. This is an application of the principle of maximum entropy [24]-[27]. The expression (1) is the entropy gain in a stochastic process that is passed through a linear filter that converts white noise to a process with power spectral density $S(f)$ (see [9, pp. 412-414], [29, pp. 93-95], [30, p. 243]). This suggests that the process entropy can be maximized by maximizing (1) subject to the constraints (2). The result is

$$S(f) = \frac{\sigma^2}{\left| \sum_{r=0}^{M} a_r z^{-r} \right|^2} \tag{8}$$

where $z = \exp(-2\pi i f \Delta t)$. This is the familiar MESA [2] or linear prediction coding (LPC) [4] estimate. Such derivations of (8) have several mathematical and logical drawbacks [31]. For example, entropy is mathematically ill-behaved for continuous densities [32, pp. 31-32]. A derivation of MESA without these drawbacks arises as a special case of minimum cross-entropy spectral analysis (MCESA) [31], and also helps to expose the difference underlying the choice of maximizing (1) or (3).

Like MESA, MCESA is an information-theoretic extrapolation of $R(t)$, but it differs from MESA in that it accounts for a prior estimate of $S(f)$ [or $R(t)$]. In deriving MCESA, we lose no generality by considering time-domain signals of the form

$$s(t) = \sum_{k=1}^{N} a_k \cos(2\pi f_k t) + b_k \sin(2\pi f_k t) \tag{9}$$

where a_k and b_k are random variables and where the f_k are frequencies [31], [33, p. 36]. Since the power at frequency f_k is $x_k \equiv \frac{1}{4}(a_k^2 + b_k^2)$, we describe the random process by a joint probability density $q^\dagger(x)$, where $x = x_1, x_2, \cdots, x_N$.

The spectral power at frequency f_k of $q^\dagger(x)$ is the expectation

$$S_k^\dagger = \int x_k q^\dagger(x)\, dx. \tag{10}$$

Let P_k be a prior estimate of $S_k^\dagger$. Then it is appropriate to assume

$$p(x) = \prod_{k=1}^{N} \frac{1}{P_k} \exp \left[-\frac{x_k}{P_k} \right] \tag{11}$$

as form for the prior estimate of the probability density $q^\dagger$ [31]. Suppose that one obtains new information about $q^\dagger$ in the form of $M + 1$ values of the autocorrelation function $R(t_r)$,

$$R_r = R(t_r) = \sum_{k=1}^{N} 2 S_k^\dagger \cos(2\pi t_r f_k) \tag{12}$$

where $r = 0, \cdots, M$ and $t_0 = 0$. Using (10) one can write this in the form of expected value constraints (4) on $q^\dagger(x)$. Given the prior (11) and these constraints, one can compute a minimum cross-entropy posterior estimate $q(x)$ of the form (8). The corresponding posterior estimate of the power spectrum is just $S_k = \int dx\, x_k q(x)$, which becomes

$$S_k = \frac{1}{\frac{1}{P_k} + \sum_{r=0}^{M} 2\beta_r \cos(2\pi t_r f_k)} \tag{13}$$

where the β_k are chosen so that the S_k satisfy the autocorrelation constraints (12) (with S_k in place of $S_k^\dagger$) [31]. If one assumes a flat prior estimate of the prior spectrum, $P_k = P$, and equal spacing of the autocorrelation lags, $t_r = r\Delta t$, (13) can be written in the form (8) [31].

The posterior probability density can be expressed in terms of the posterior spectral power estimates (13)

$$q(x) = \prod_{k=1}^{N} \frac{1}{S_k} \exp \left[-\frac{x_k}{S_k} \right]. \tag{14}$$

Computing the normalized differential entropy of the posterior power estimates (13) yields

$$-\frac{1}{N} \int q(x) \log q(x) dx = 1 + \frac{1}{N} \sum_{k=1}^{N} \log S_k. \tag{15}$$

Except for the constant, which has no effect on maximization, the right-hand side of (15) is the discrete form of (1). Maximizing (15) subject to the autocorrelation constraints leads again to (8).

B. The $-S \log S$ Form

In this case we treat the unknown power spectrum variables $S_k^\dagger$ as probabilities, which is mathematically reasonable provided that the power spectrum is normalized so that $\Sigma_k S_k^\dagger = 1$. The known autocorrelations are then expressed as expectations of the probability distribution $S_k^\dagger$, $k = 1, \cdots, N$, as in (12).

In deriving the resulting power spectrum estimate, we again proceed with the general case involving a prior estimate and cross-entropy minimization. Since we assumed a known autocorrelation for lag $t_0 = 0$, $\Sigma_k S_k^\dagger = \frac{1}{2} R_0$ is known. Let P_k be a prior estimate of $S_k^\dagger$, and let $\boldsymbol{q}^\dagger = \{q_1^\dagger, q_2^\dagger, \cdots, q_N^\dagger\}$ and $\boldsymbol{p} = \{p_1, p_2, \cdots, p_N\}$ be probability distributions defined by normalizing the power spectra $S_k^\dagger$ and P_k, i.e., $q_k^\dagger = (2S_k^\dagger)/R_0$ and $p_k = P_k/T$, where $T = \Sigma_k P_k$. We rewrite the autocorrelation constraints (12) as expectations of $\boldsymbol{q}^\dagger$:

$$\frac{2R_r}{R_0} = \sum_{k=1}^{N} 2\cos(2\pi t_r f_k) q_k^\dagger \tag{16}$$

and we obtain a posterior estimate $\boldsymbol{q}$ of $\boldsymbol{q}^\dagger$ by minimizing the cross-entropy $H(\boldsymbol{q}, \boldsymbol{p})$ subject to the constraints (16) (with q_k in place of $q_k^\dagger$). Note that the constraint for $r = 0$ reduces to the normalization constraint $\Sigma_k q_k = 1$. The result is

$$q_k = p_k \exp\left[-\sum_{r=0}^{M} 2\mu_r \cos(2\pi t_r f_k)\right] \tag{17}$$

where the μ_r are chosen to satisfy the constraints. We define the posterior power spectrum estimate as $S_k = \frac{1}{2}R_0 q_k$, which satisfies (12). This yields

$$S_k = P_k \exp\left[-\sum_{r=0}^{M} 2\varphi_r \cos(2\pi t_r f_k)\right] \tag{18}$$

where the φ_r are equal to the μ_r in (17) except for φ_0, which satisfies $\varphi_0 = \mu_0 + \frac{1}{2}\log(R_0/2T)$. Since

$$\sum_{k=1}^{N} S_k \log\frac{S_k}{P_k} = \tfrac{1}{2}R_0 H(\boldsymbol{q}, \boldsymbol{p}) + \tfrac{1}{2}R_0 \log\frac{R_0}{2T}$$

it follows that minimizing $H(\boldsymbol{q}, \boldsymbol{p})$ is equivalent to minimizing

$$\sum_{k=1}^{N} S_k \log\frac{S_k}{P_k} \tag{19}$$

so that minimizing (19) subject to the constraints (12) yields (18). For a flat prior estimate $P_k = P$, minimizing (19) is equivalent to maximizing

$$-\sum_{k=0}^{N} S_k \log S_k$$

which is just the discrete form of (3).

C. Discussion

Both estimates of $S_k^\dagger$ proceed from a prior estimate P_k and known autocorrelations R_r. When the coefficients in an underlying Fourier series model are treated as random variables and the $S_k^\dagger$ are treated as expectations, cross-entropy minimization leads to the estimate (13). For the case of a flat prior estimate $P_k = P$, (13) follows from maximizing $\Sigma_k \log S_k$. When the $S_k^\dagger$ are treated as probabilities rather than expectations, cross-entropy minimization leads to the estimate (18), which also follows from maximizing $-\Sigma_k S_k \log S_k$ in the case of a flat prior. Because the result in this case arises from performing maximum entropy on a probability distribution defined by normalizing a power spectrum, we refer to it as maximum entropy normalized spectral analysis (MENSA).[1] The Lagrangian multipliers β_r in (13) and φ_r in (18) are chosen in both cases so that the estimates agree with the known autocorrelations

$$R_r = \sum_{k=1}^{N} 2S_k \cos(2\pi t_r f_k) \qquad (r = 0, 1, \cdots, M). \tag{20}$$

Given one of the spectral estimates S_k, $k = 1, \cdots, N$, substitution of an arbitrary lag t for t_r in (20) defines the corresponding extrapolation of the known autocorrelations.

Which of the two estimates (13) and (18) is better? In our opinion, if one has a good physical model for some variable of interest, and if the model can be incorporated into the derivation of an estimate for that variable, it makes sense to do so. Because such estimates can exploit more information than estimates derived without an underlying model, estimates based on underlying models should be better. Since (13) exploits an underlying Fourier series model—well known to be a useful model for time series—this point favors (13). Also, since (13) yields all-pole models in the important case of flat priors, since all-pole spectra result from passing a broad-band signal through a multilayered transmission medium, and since the human vocal tract is a multilayered transmission medium, it follows that (13) should be appropriate for speech processing.

On the other hand, arguments for (18) also have merit. For example, in arguing for the maximization of $-\Sigma_k S_k \log S_k$ rather than $\Sigma_k \log S_k$, Skilling [16] points out that the goal is to estimate the power spectrum itself, not the Fourier amplitudes in an underlying model like (9), so that a more direct and better estimate should result from treating the unknown power spectrum variables $S_k^\dagger$ as probabilities. Mathematically, this is reasonable provided that the power spectrum is normalized so that $\Sigma_k S_k^\dagger = 1$. Furthermore, speech spectra are known to have occasional zeros, and the form of (18) shows that small values for S_k can arise from moderate values of the trigonometric polynomial in the exponent. The MESA estimate is well known to have difficulty estimating zeros. An additional reason to consider (18) seriously is its success in other fields, which we have already mentioned. Furthermore, spectral estimates based on the minimization of (19) have been reported recently in [34] and a first-order approximation of the estimate (18) appears to be equivalent to the PDFT estimator introduced in [35], [36].

These arguments do not clearly favor one estimator or the other. While the success of (13) in speech processing is strong evidence, it seems clear that the potential of (18) will continue to be raised, so that an empirical evaluation is necessary. This we attempt to do in the remainder of this paper.

III. Experimental Approach

This section contains basic definitions, a discussion of our experimental approach, and a discussion of various computational issues. Our general approach is to process various speech

[1] This somewhat contrived acronym has the additional virtue of being the Latin word for "table," which is the source of the Spanish word for table (*mesa*).

signals in order to compare measured power spectra and autocorrelations with MESA and MENSA estimates. We also synthesize speech using both MESA and MENSA power spectrum estimates and perform qualitative comparisons of the results.

A. Definitions and Notation

Let $\boldsymbol{y} \equiv \{y_1, y_2, \cdots, y_L\}$ comprise L time-domain samples, equispaced at intervals of Δt, from one "frame" of speech data. From $\boldsymbol{y}$, we compute estimated autocorrelations $\boldsymbol{R} \equiv \{R_0, R_1, \cdots, R_{L-1}\}$ by means of

$$R_r = \frac{1}{L} \sum_{i=1}^{L-r} y_i y_{i+r}. \tag{21}$$

This is a biased estimate but it guarantees positive-definiteness. Let $\boldsymbol{Q} \equiv \{Q_1, Q_2, \cdots, Q_N\}$ be the power spectrum defined by the discrete Fourier transform of the measured autocorrelations,

$$Q_k = R_0 + \sum_{r=1}^{L-1} 2R_r \cos(2\pi t_r f_k). \tag{22}$$

As the N discrete frequencies we take $f_k = (k - \frac{1}{2})/(2N\Delta t)$.

Let $\boldsymbol{S} \equiv \{S_1, S_2, \cdots, S_N\}$ be the power spectrum estimate obtained from (13) using a flat prior estimate and the first $M + 1$ autocorrelations R_r from (21). $\boldsymbol{S}$ is the standard MESA or LPC estimate of the power spectrum—its usual, continuous-frequency form is given by (8). Let $\boldsymbol{S}^* \equiv \{S_1^*, S_2^*, \cdots, S_N^*\}$ be the MENSA power spectrum estimate obtained from (18) using the same flat prior and the same $M + 1$ autocorrelations. Finally, let $\boldsymbol{A}$ and $\boldsymbol{A}^*$ be the extrapolated autocorrelations for all L lags $t_r = r\Delta t$, $r = 0, \cdots, L - 1$, obtained from (20) using $\boldsymbol{S}$ and $\boldsymbol{S}^*$, respectively. Note that A_r and A_r^* match the actual autocorrelations (21) for $r = 0, \cdots, M$. For $r > M$, however, they are in general different from each other and from R_r. For convenience, we summarize the notation as follows:

$\boldsymbol{y}$ vector of L time-domain samples from one speech frame
$\boldsymbol{R}$ the measured autocorrelations for L lags computed from $\boldsymbol{y}$
$\boldsymbol{Q}$ "actual" power spectrum defined by a Fourier transform of $\boldsymbol{R}$
$\boldsymbol{S}$ MESA or LPC estimate of power spectrum from first $M + 1$ lags of $\boldsymbol{R}$
$\boldsymbol{S}^*$ MENSA estimate of power spectrum from first $M + 1$ lags of $\boldsymbol{R}$
$\boldsymbol{A}$ MESA or LPC autocorrelation extrapolation based on $\boldsymbol{S}$
$\boldsymbol{A}^*$ MENSA autocorrelation extrapolation based on $\boldsymbol{S}^*$.

For the work reported here, we used $L = 180$ and $M = 8,10,25$. When we refer to more than one speech frame, we add a subscript to the foregoing definitions.

B. What and How to Compare

In order to compare MESA and MENSA, we did three things. 1) For a variety of representative speech frames, we plotted $\boldsymbol{A}$, $\boldsymbol{A}^*$, and $\boldsymbol{R}$ and compared them. 2) For the same frames, we plotted $\boldsymbol{S}$, $\boldsymbol{S}^*$, and $\boldsymbol{Q}$ and compared them. 3) We compared speech synthesized two different ways: we used identical pitch and voicing decisions, and either $\boldsymbol{S}$ or $\boldsymbol{S}^*$ for spectral shape. All of these comparisons were qualitative.

What about quantitative comparisons? For some distortion measure d, one could compare $d(\boldsymbol{Q}, \boldsymbol{S})$ with $d(\boldsymbol{Q}, \boldsymbol{S}^*)$, but what should d be? One distortion measure could yield $d(\boldsymbol{Q}, \boldsymbol{S}) < d(\boldsymbol{Q}, \boldsymbol{S}^*)$ while another could yield the reverse inequality. One reasonable choice is the Itakura-Saito distortion d_{IS} [37],

$$d_{\text{IS}}(\boldsymbol{Q}, \boldsymbol{S}) = \frac{1}{N} \sum_{k=1}^{N} \left[\frac{Q_k}{S_k} - 1 - \log\frac{Q_k}{S_k}\right]$$

which is known to be useful in speech processing. But, in the notation of Section II, the Itakura-Saito distortion $d_{\text{IS}}(\boldsymbol{S}, \boldsymbol{P})$ is just the asymptotic cross-entropy $H(q, p)$—derivations of MESA spectra by cross-entropy minimization are equivalent to derivations by minimization of d_{IS} [10], [31], [38]. Not only does $\boldsymbol{S}$ minimize $d_{\text{IS}}(\boldsymbol{S}, \boldsymbol{P})$ subject to the constraints, but $\boldsymbol{S}$ is the spectrum of the form (13) that minimizes $d_{\text{IS}}(\boldsymbol{Q}, \boldsymbol{S})$ [37], [39]. Use of d_{IS} might therefore involve an intrinsic bias in favor of MESA. We therefore consider a distortion measure that bears a relation to MENSA analogous to that of d_{IS} to MESA. Define the "cross-entropy distortion" $d_{\text{CE}}(\boldsymbol{Q}, \boldsymbol{S})$ to be the cross entropy of the probability distributions obtained by normalizing $\boldsymbol{Q}$ and $\boldsymbol{S}$

$$d_{\text{CE}}(\boldsymbol{Q}, \boldsymbol{S}) = \sum_{k=1}^{N} \frac{Q_k}{\sum_{j=1}^{N} Q_j} \log\frac{Q_k}{S_k} - \log \frac{\sum_{j=1}^{N} Q_j}{\sum_{j=1}^{N} S_j}.$$

Then $\boldsymbol{S}^*$ minimizes $d_{\text{CE}}(\boldsymbol{S}^*, \boldsymbol{P})$ subject to constraints just as $\boldsymbol{S}$ minimizes $d_{\text{IS}}(\boldsymbol{S}, \boldsymbol{P})$ subject to constraints. Moreover, $\boldsymbol{S}^*$ is one of the spectra of the form (21) that minimizes $d_{\text{CE}}(\boldsymbol{Q}, \boldsymbol{S}^*)$ [22]. We also use a third distortion measure, the gain-optimized Itakura-Saito distortion defined by $d_{\text{GO}}(\boldsymbol{Q}, \boldsymbol{S}) = \min_g d_{\text{IS}}(g\boldsymbol{Q}, \boldsymbol{S})$, where g ranges over positive constant scale factors [39]. This is closely related to d_{IS} but, like d_{CE}, is insensitive to changes in the gains of the two spectra. It can be computed from the formula

$$d_{\text{GO}}(\boldsymbol{Q}, \boldsymbol{S}) = \log\frac{1}{N} \sum_{k=1}^{N} \frac{Q_k}{S_k} - \frac{1}{N} \sum_{k=1}^{N} \log\frac{Q_k}{S_k}.$$

C. Numerical Issues and Procedures

The MENSA estimate $\boldsymbol{S}^*$ can be produced by an algorithm that determines minimum cross-entropy probability distributions given arbitrary priors and arbitrary constraints [22], [40]. For the work reported here, we used a Fortran version of the Newton-Raphson based APL program described in [40]. The resulting spectrum $\boldsymbol{S}^*$ may be thought of as a discrete-frequency approximation to a continuous power spectrum. Clearly, the accuracy of the discrete-frequency approximation will depend on the number of frequency points N.

As for $\boldsymbol{S}$, a variety of methods are available. Standard MESA or LPC methods can produce the a_r used in (8) or any of the equivalent sets of parameters such as reflection coefficients. The result is a continuous representation of the spectrum estimate that can then be evaluated at the frequencies f_k in order to yield $\boldsymbol{S}$. This is more accurate than methods that compute discrete-frequency approximations, but to use it might introduce a misleading source of differences between $\boldsymbol{S}$ and $\boldsymbol{S}^*$. We therefore chose to compute the $\boldsymbol{S}$ in a manner analogous to the computation of $\boldsymbol{S}^*$. In particular, we used a Fortran implementation of the MCESA [31] algorithm described in [41], which uses the Newton-Raphson method to compute (13) for arbitrary priors and autocorrelation constraints. For a flat prior, the result is just a discrete-frequency approximation to a continuous MESA or LPC spectrum. As checks on the discrete-frequency computations of $\boldsymbol{S}^*$ and $\boldsymbol{S}$, we obtained results for various values of the number of frequencies N, and we compared the results for $\boldsymbol{S}$ with continuous frequency results obtained using Levinson recursion.

To obtain synthetic speech using the two different spectral shapes, we used commonly-available, LPC-based programs. Our procedure was as follows. First we analyzed the test sentence for pitch and voicing using a modified cepstral technique described in [42] and implemented in Version 4.0 of the Interactive Laboratory System (ILS) from Signal Technology, Inc. The results were used for both syntheses. For the synthesis based on $\boldsymbol{S}^*$, we used a 29th-order all-pole approximation to the power spectrum $\boldsymbol{S}^*$ in each frame. This approximation was computed by taking the first 29 lags of the autocorrelation extrapolation $\boldsymbol{A}^*$ and using Levinson recursion to yield a set of reflection coefficients. As checks, we plotted the resulting approximate power spectrum and compared it with $\boldsymbol{S}^*$. For the synthesis based on $\boldsymbol{S}$, we followed the analogous procedure—we ran Levinson recursion on the first 29 lags of $\boldsymbol{A}$. Had we been dealing with exact, continuous spectra, the resulting "approximate" spectrum would be exactly equal to $\boldsymbol{S}$, so it would have been reasonable to bypass this step. We included it, however, in order to keep the comparison as fair as possible. As a check, we also synthesized speech using spectra obtained directly from Levinson recursion on the first $M + 1$ lags of the measured autocorrelations $\boldsymbol{R}$. Note that the 29th-order all-pole synthesis spectra are 29th-order approximations to $\boldsymbol{S}$ and $\boldsymbol{S}^*$, and not 29th-order approximations to $\boldsymbol{Q}$.

IV. Experimental Results

We obtained results for the sentence "The meeting begins at four P.M." The sentence was spoken by a male, passed through an antialiasing filter, digitized at 8000 samples/s, and divided into 100 frames of 180 samples each. Using 256 discrete frequencies ($N = 256$), we computed $\boldsymbol{R}_j$, $\boldsymbol{Q}_j$, $\boldsymbol{S}_j^*$, $\boldsymbol{A}_j^*$, $\boldsymbol{S}_j$, and $\boldsymbol{A}_j$, $j = 1, \cdots, 100$, as discussed in the previous section. We also did computations for some cases with $N = 64$ and N = 128. In general, there were no essential differences between results for $N = 64$ or 128 and $N = 256$. We also repeated the computations using Hamming windowing alone, 90 percent preemphasis alone, and both together on the digitized speech. In the following, we focus attention on two frames—frame 56, which contains a portion of the phoneme /f/, and frame 39, which contains a portion of the phoneme /ɪ/. We refer to these frames by means of the subscripts f and I, respectively. Unless windowing or preemphasis is explicitly mentioned, the reference is to the spectra computed without preprocessing.

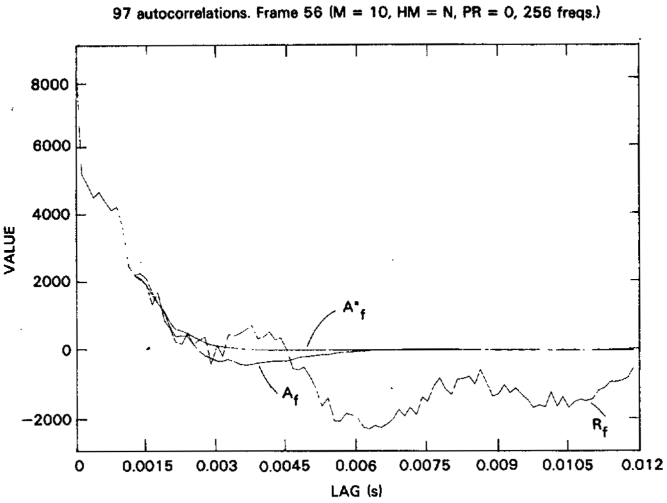


Fig. 1. Autocorrelations from speech samples and from MESA and MENSA spectral estimates (/f/).

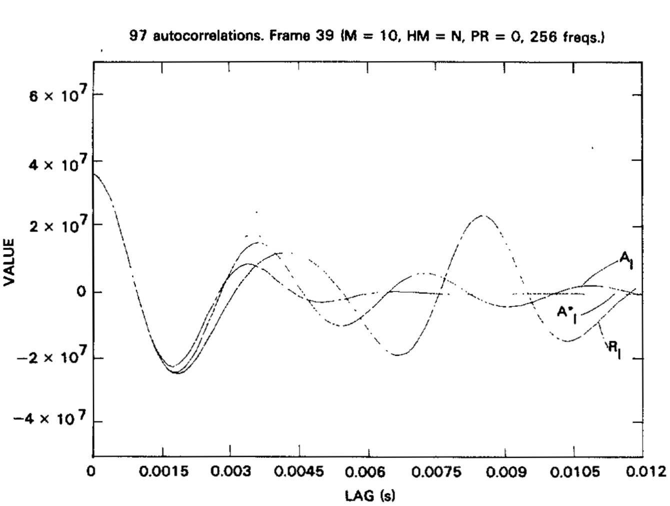


Fig. 2. Autocorrelations from speech samples and from MESA and MENSA spectral estimates (/ɪ/).

A. Comparison of Autocorrelation Extrapolations

In Fig. 1, we plot $\boldsymbol{R}_f$, $\boldsymbol{A}_f^*$, and $\boldsymbol{A}_f$ for $N = 256$. When we plotted the continuous autocorrelation function obtained by Levinson recursion, it was indistinguishable from $\boldsymbol{A}_f$, which implies that the discrete frequency approximations are accurate. Beyond the constraint limit of lag 10, the extrapolations $\boldsymbol{A}_f^*$ and $\boldsymbol{A}_f$ differ from each other as well as from $\boldsymbol{R}$. One would be hard pressed to argue that either one is a "better" extrapolation. The same conclusion follows from Fig. 2, in which we plot $\boldsymbol{R}_I$, $\boldsymbol{A}_I^*$, and $\boldsymbol{A}_I$.

B. Comparison of Power Spectra

Turning to the power spectra, we plot $\boldsymbol{S}_f^*$, $\boldsymbol{S}_f$, $\boldsymbol{S}_I^*$, and $\boldsymbol{S}_I$ in Figs. 3-6 for $M = 10$. The spectra $\boldsymbol{S}_f^*$ and $\boldsymbol{S}_f$ are quite similar;

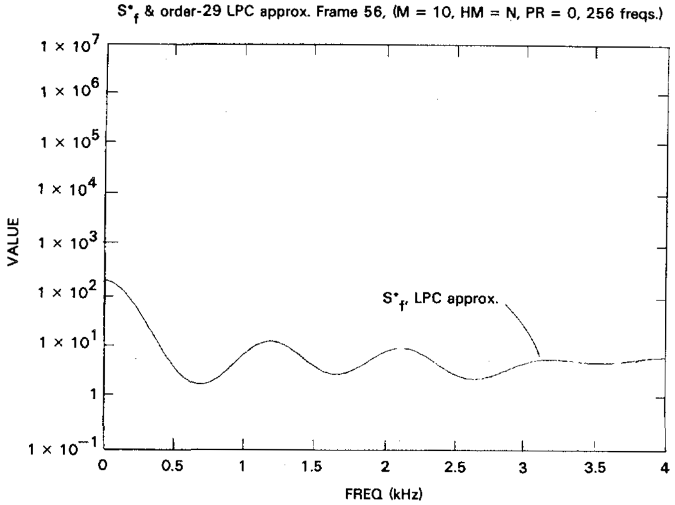


Fig. 3. MENSA spectrum and 29th-order continuous MESA approximation (/f/).

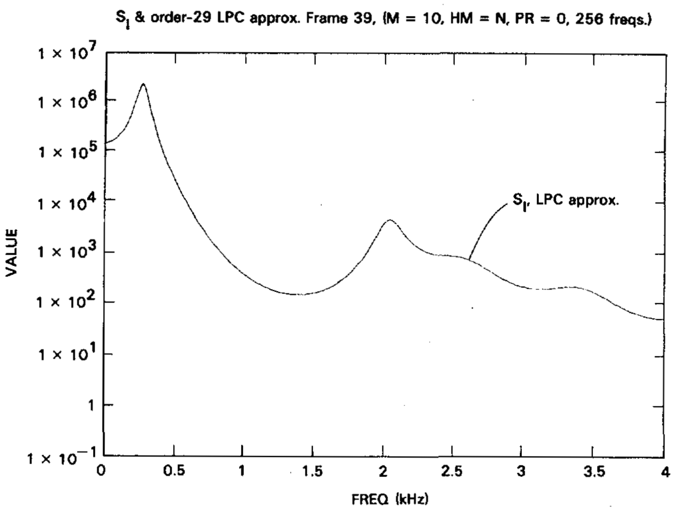


Fig. 6. Discrete MESA spectrum and 29th-order continuous MESA approximation (/ɪ/).

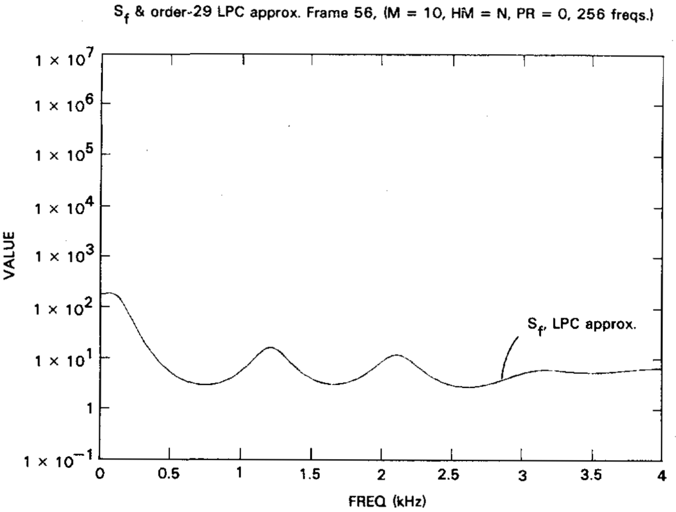


Fig. 4. Discrete MESA spectrum and 29th-order continuous MESA approximation (/f/).

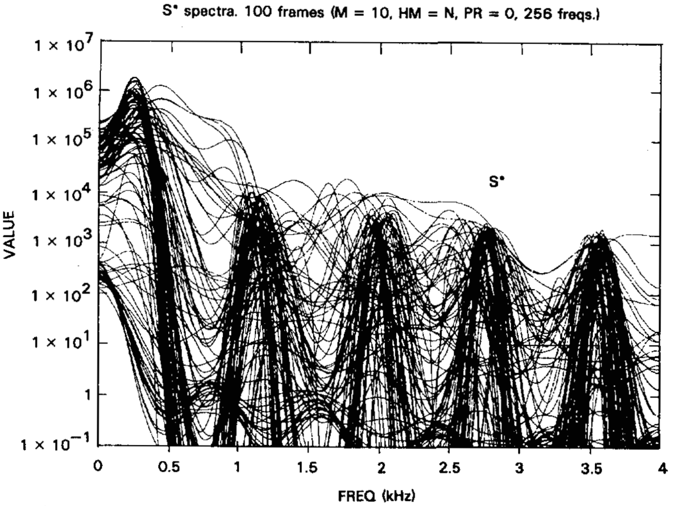


Fig. 7. MENSA spectra–100 frames overlaid.

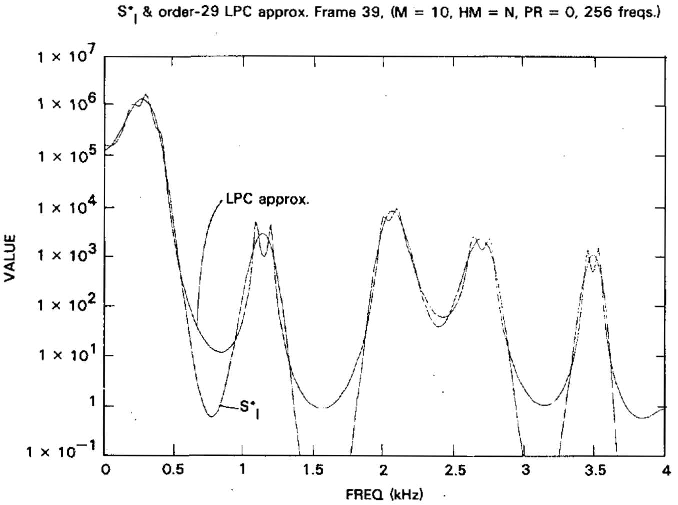


Fig. 5. MENSA spectrum and 29th-order continuous MESA approximation (/ɪ/).

S_I^* and S_I are quite different. In particular, S_I^* has deep nulls that are characteristic of the MENSA estimates for the whole test sentence. Indeed, the frequent occurrence of five lobes is obvious in Fig. 7, which shows the superimposed results of S^* for all 100 frames ($N = 256$), confirming an earlier conjecture about spectral zeros. No such structure occurs for S (Fig. 8). The lobe structure appears to be related to the number of constraints: there are five lobes in Fig. 7, which is one half the analysis order ($M = 10$). We repeated the computation of A^* using $M = 25$ and $M = 8$. The resulting plots were similar to Fig. 7 except that about 12 and 4 lobes were apparent, respectively. Neither preemphasis nor windowing was entirely effective in eliminating the deep minima from the MENSA spectra. The superposed plots continued to show a lobed structure, although more complex and less regular than the consistent five-lobe pattern of Fig. 7. The results of using both Hamming windowing and 90 percent preemphasis are shown in Fig. 9.

In Fig. 10, we compare the "actual" power spectrum Q_f with S_f^* and S_f. Both estimates appear to be smoothed versions of Q_f. Fig. 11 shows the analogous comparison for /ɪ/. Here

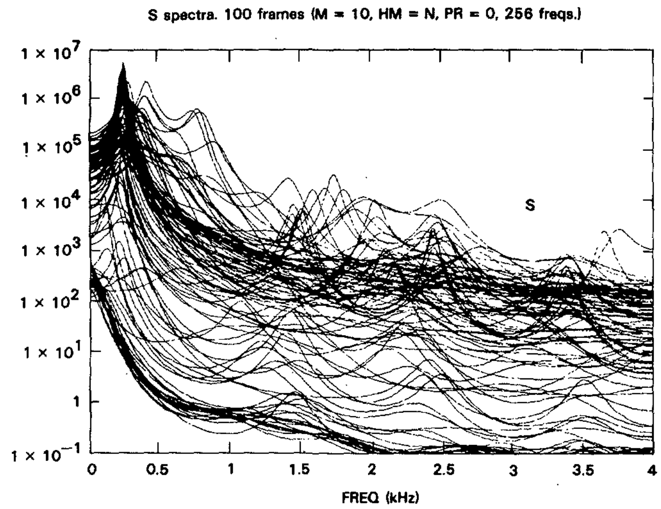


Fig. 8. MESA spectra–100 frames overlaid.

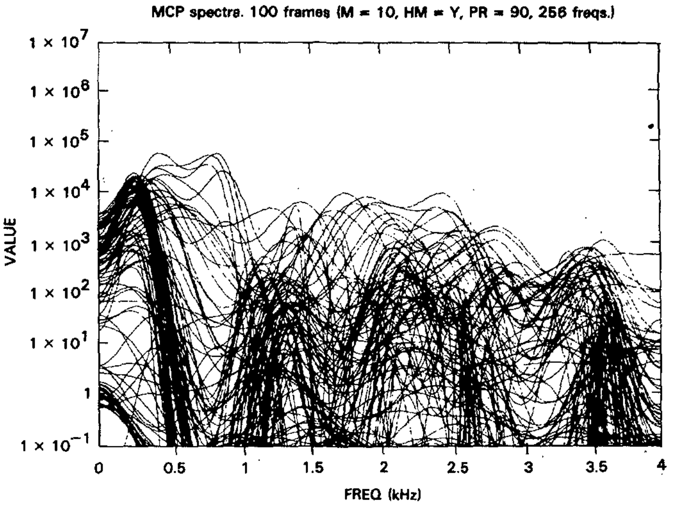


Fig. 9. MENSA spectra from windowed, preemphasized speech–100 frames overlaid.

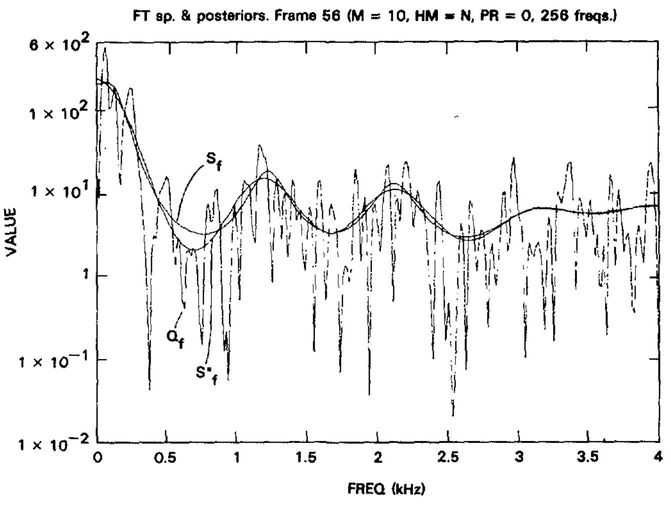


Fig. 10. MESA and MENSA estimates with Fourier transform of measured autocorrelations (/f/).

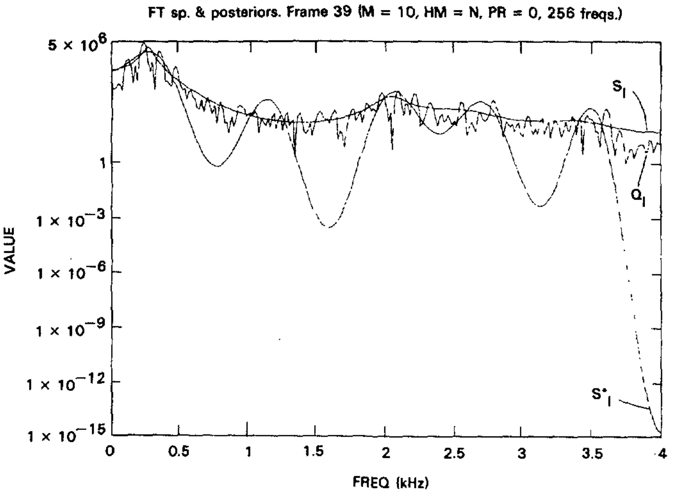


Fig. 11. MESA and MENSA estimates with Fourier transforms of measured autocorrelations (/ɪ/).

TABLE I
DISTORTION RESULTS

			Itakura-Saito		Gain-Opt. I-S		Cross-Entropy	
M	Pr.	Win.	MESA	MENSA	MESA	MENSA	MESA	MENSA
8	0	none	0.320	1.6×10^{20}	0.320	9.970	0.361	0.570
10	0	none	0.275	3.7×10^{19}	0.275	10.549	0.307	0.495
25	0	none	0.204	5.8×10^{18}	0.204	4.352	0.165	0.290
8	90%	Ham.	0.589	6.1×10^{18}	0.589	12.340	0.343	0.507
10	90%	Ham.	0.502	3.3×10^{18}	0.502	11.025	0.292	0.429
25	90%	Ham.	0.310	2.5×10^{17}	0.310	4.776	0.167	0.162
8	90%	none	0.434	8.5×10^{15}	0.434	5.520	0.428	0.596
10	90%	none	0.379	1.1×10^{19}	0.379	5.019	0.359	0.521
25	90%	none	0.285	1.1×10^{2}	0.285	1.032	0.194	0.282
8	0	Ham.	0.553	1.6×10^{20}	0.553	15.265	0.354	0.374
10	0	Ham.	0.446	8.7×10^{19}	0.446	16.347	0.243	0.321
25	0	Ham.	0.290	4.2×10^{19}	0.290	13.203	0.134	0.164

there is more of a difference, and it appears more reasonable to interpret S_I than S_I^* as a smoothed version of Q_I.

For three values of M, we computed six distortion measures between Q and the estimates S and S^*: $d_{\text{IS}}(Q, S)$, $d_{\text{IS}}(Q, S^*)$, $d_{\text{GO}}(Q, S)$, $d_{\text{GO}}(Q, S^*)$, $d_{\text{CE}}(Q, S)$, and $d_{\text{CE}}(Q, S^*)$. The results, averaged over all 100 frames, are shown in Table I. In one case the mean distortion for MENSA is slightly less than that for MESA, the difference being in the third decimal place. In every other case the mean distortion for MESA is less. This is true even for the "cross-entropy" distortions d_{CE}, which might have been expected to favor MENSA. The d_{CE} results do not favor MESA as overwhelmingly as those from the other two distortion measures–especially d_{IS}. The enormous values of d_{IS} for the MENSA spectra are the result of the deep minima. The other two distortion measures contain the term Q_k/S_k^* only logarithmically. Thus, d_{IS} penalizes underestimates more severely than do d_{GO} and d_{CE}.

Two columns of the table are identical: it appears that $d_{\text{IS}}(Q, S) = d_{\text{GO}}(Q, S)$. This is no coincidence, but a property of d_{IS} and d_{GO}. The equality can be shown to hold provided that S is a MESA spectrum and that Q is a spectrum that satisfies the same autocorrelation constraints that determine S. A proof can be based on the "correlation matching" property [22], [39] of MESA spectra.

C. Comparison of Synthetic Speech

Although results such as Fig. 7, Fig. 11, and Table I suggest that S is better than S^*, they are hardly compelling. This is a case where the proof must be in the hearing. Consequently, we synthesized the entire test sentence using the 29th-order LPC approximations as discussed in Section III-C. The 29th-order LPC approximations to S_f^*, S_f, S_I^*, and S_I are also plotted in Figs. 3-6. The two curves are indistinguishable in Figs. 3, 4, and 6; the only discrepancy is for S_I^* (Fig. 5). In that case, the 29th-order approximation is unable to match the deep nulls and also exhibits some peak splitting.

The standard LPC speech and the speech based on S sounded identical, adding further confidence to the discrete frequency approximations. The versions based on S and S^* sounded different, but—somewhat to our surprise—we and others judged them to be equally intelligible. There was, however, a distinct qualitative difference when preemphasis was not used. The speech based on S^* was qualitatively inferior—it had a distinct ringing quality, as though spoken from the other end of a long, wide pipe. When preemphasis was used, alone or with Hamming windowing, the ringing quality was greatly reduced or effectively eliminated. Hamming windowing alone reduced the ringing only slightly. We hypothesize that this ringing effect is a reflection of the characteristic lobe structure and deep minima of the spectral estimates S^*, since the ringing is most prominent when the lobing is most prominent and regular. However, the ringing can be almost imperceptible while lobing is still plainly visible in spectral plots.

V. Conclusions

Based primarily on the results of speech synthesis, but also on results like Fig. 7, Fig. 11, and Table I, we believe that it is fair to conclude that MESA (S) yields better power spectrum estimates for speech processing than does MENSA (S^*).

References

[1] J. P. Burg, "Maximum entropy spectral analysis," presented at 37th Annu. Meet. Soc. Explor. Geophys., Oklahoma City, OK, 1967.

[2] —, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1975 (University Microfilms no. 75-25,499).

[3] A. Van Den Bos, "Alternative interpretation of maximum entropy spectral analysis," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 493-494, July 1971.

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

[5] S. M. Kay and S. L. Marple, Jr., "Spectrum analysis—A modern perspective," *Proc. IEEE*, vol. 69, pp. 1380-1419, Nov. 1981.

[6] A. Papoulis, "Maximum entropy and spectral estimation: A review," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1176-1186, Dec. 1981.

[7] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, pp. 661-675, 1971.

[8] T. J. Ulrych and T. N. Bishop, "Maximum entropy spectral analysis and autoregressive decomposition," *Rev. Geophys. Space Phys.*, vol. 43, pp. 183-200, 1975.

[9] D. E. Smylie, G. K. C. Clarke, and T. J. Ulrych, "Analysis of irregularities in the earth's rotation," in *Methods in Computational Physics*, vol. 13. New York: Academic, 1973, pp. 391-431.

[10] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information measure," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 708-721, Nov. 1981.

[11] S. J. Wernecke and L. D'Addario, "Maximum entropy image reconstruction," *IEEE Trans. Comput.*, vol. C-26, pp. 351-364, Apr. 1977.

[12] J. G. Ables, "Maximum entropy spectral analysis," *Astron. Astrophys. Suppl.*, vol. 15, pp. 383-393, 1974.

[13] S. J. Wernecke, "Two-dimensional maximum entropy reconstruction of radio brightness," *Radio Sci.*, vol. 12, no. 5, pp. 831-844, 1977.

[14] B. R. Frieden, "Restoring with maximum likelihood and maximum entropy," *J. Opt. Soc. Amer.*, vol. 62, pp. 511-518, Apr. 1972.

[15] R. Gordon and G. T. Herman, "Reconstruction of pictures from their projections," *Quart. Bull. Cen. Theor. Biol.*, vol. 4, pp. 71-151, 1971.

[16] J. Skilling, "Maximum entropy and image processing—Algorithms and applications." in *Proc. 1st Maximum Entropy Workshop*, 1981.

[17] M. D. Ortigueira, R. Garcia-Gomez, and J. M. Tribolet, " An iterative algorithm for maximum flatness spectral analysis," in *Proc. Int. Conf. DSP*, 1981, pp. 810-818.

[18] C. Nadeu, E. Sanvicente, and M. Bertran, "A new algorithm for spectral estimation," in *Proc. Int. Conf. DSP*, 1981, pp. 463-470.

[19] R. Kikuchi and B. H. Soffer, "Maximum entropy image restoration. I. The entropy expression," *J. Opt. Soc. Amer.*, vol. 67, no. 12, pp. 1656-1665, 1977.

[20] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959; New York: Dover (reprint), 1969.

[21] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26-37, Jan. 1980; see also comments and corrections, *IEEE Trans. Inform. Theory*, vol. IT-29, Nov. 1983.

[22] —, "Properties of cross-entropy minimization," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 472-482, July 1981.

[23] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 227-241, 1968.

[24] M. R. Schroeder, "Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis," *Speech Commun.*, vol. 1, pp. 9-20, 1982.

[25] E. T. Jaynes, "Information theory and statistical mechanics I," *Phys. Rev.*, vol. 106, pp. 620-630, 1957.

[26] —, "Information theory and statistical mechanics II," *Phys. Rev.*, vol. 108, pp. 171-190, 1957.

[27] W. M. Elsasser, "On quantum measurements and the role of the uncertainty relations in statistical mechanics," *Phys. Rev.*, vol. 52, pp. 987-999, Nov. 1937.

[28] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Math. Statist.*, vol. 3, pp. 146-158, 1975.

[29] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Chicago, IL: Univ. Illinois Press, 1949.

[30] M. S. Bartlett, *An Introduction to Stochastic Processes.* Cambridge, England: Cambridge Univ. Press, 1966.

[31] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 230-237, Apr. 1981.

[32] R. G. Gallagher, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[33] Y. Yaglom, *An Introduction to the Theory of Stationary Random Functions.* Englewood Cliffs, NJ: Prentice-Hall, 1962.

[34] L. Vergara-Domínguez and A. R. Figueiras-Vidal, "A minimum cross-flatness spectral estimator and some related problems," preprint, private communication.

[35] C. L. Byrne and R. M. Fitzgerald, "Reconstruction from partial information with applications to tomography," *SIAM J. Appl. Math.*, vol. 42, pp. 933-940, Aug. 1982.

[36] R. M. Fitzgerald, private communication.

[37] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Rep. 6th Int. Cong. Acoust.*, 1968.

[38] —, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53-A, pp. 36-43, 1970.

[39] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367-376, Aug. 1980.

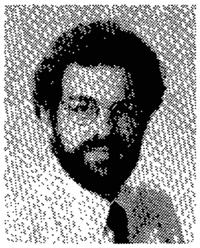
[40] R. W. Johnson, "Determining probability distributions by maximum entropy and minimum cross-entropy," in *APL79 Conf. Proc.*, ACM Pub. no. 0-89791-005, May 1979.

[41] —, "Algorithms for single-signal and multisignal minimum-cross-entropy spectral analysis," Naval Res. Lab., Washington, DC, NRL Rep. 8667.

[42] R. W. Schafer and J. D. Markel, Eds., *Speech Analysis.* New York: IEEE Press, 1979.

Rodney W. Johnson was born in Scottsbluff, NE, on March 8, 1940. He received the B.S. degree in physics and the Ph.D. degree in mathematics from Stanford University, Stanford, CA, in 1962 and 1967, respectively.

He held research associateships in physics at Princeton University, Princeton, NJ, from 1966 to 1968 and at Syracuse University, Syracuse, NY, from 1968 to 1970. From 1970 to 1973 he was an Assistant Professor in the Department of Mathematics Fordham University, New York, NY. Following a year at the Center for Naval Analyses, he joined the Naval Research Laboratory, Washington, DC, in 1974, where he is a member of the Computer Science Section in the Computer Science and Systems Branch of the Information Technology Division. His current and former interests include information theory, signal processing, speech synthesis, and mathematical physics.



John E. Shore (M'72-SM'81) was born in England on September 2, 1946. He received the B.S. degree in physics from Yale University, New Haven, CT, in 1968, and the Ph.D. degree in theoretical physics (statistical mechanics) from the University of Maryland, College Park, in 1974.

In 1968 he joined the Naval Research Laboratory, Washington, DC, where he is currently Head of the Computer Science Section in the Computer Science and Systems Branch of the Information Technology Division. His previous research interests include computer architecture, dynamic memory allocation, programming language design, software engineering, and text-to-speech translation. His current interests include information theory, queuing theory, system modeling, pattern recognition, spectrum analysis, and speech processing.

Design of Antialiasing Patterns for Time-Sequential Sampling of Spatiotemporal Signals

JAN P. ALLEBACH

Abstract—**The aliasing that results from time-sequential sampling of spatiotemporal signals is strongly dependent on the order in which the spatial points are sampled. To design sampling patterns that reduce aliasing, the sequence of sampling points is mapped into several shorter subsequences via the chinese remainder theorem. A pairwise exchange algorithm then finds the best ordering of each subsequence. The patterns obtained with this procedure perform substantially better than those known previously, and perform as well as the optimal patterns that can be expressed in closed form when the signal is termporally undersampled by less than a factor of 2.**

I. Introduction

MANY signal processing and communications problems involve time-varying images. To be processed digitally, these signals must be sampled in space and time. It is common practice to do the sampling in a time-sequential fashion, collecting a frame of samples one-by-one from the spatial region and then repeating this process. The scanning action may be generated electromechanically or by multiplexing the outputs from an array of sensors.

With some systems such as sensor arrays, the spatial points may be sampled in any order. With other systems, the ordering may be partially constrained by the scanning mechanism. In either case, the points are most frequently taken in lexicographic order which in 2 spatial dimensions results in line-by-line scanning. A number of researchers have experimented with other orderings of the spatial points [1]-[5]. In particular, Deutsch [1], [3] proposed an ordering which tends to distribute the samples taken during any time interval of duration less than the frame period uniformly over the spatial region. Since the ordering may be generated by a mapping from the bit reversed output of a binary counter, we refer to it as the bit reversed sampling pattern.

Fig. 1 shows the lexicographic and bit reversed sampling patterns for one spatial dimension. During each frame period of duration B, M samples are taken uniformly over the spatial region at interval X. With either pattern, we would expect to resolve signal components with temporal frequency $f_0 < 1/(2B)$ and spatial frequency $u_0 < 1/(2X)$. With the bit reversed pattern samples taken during a time interval B/λ, $1 \leqslant \lambda \leqslant M$ are distributed at approximately a spatial interval of λX. With this pattern, we might expect to also resolve signal components with $f_0 < \lambda/(2B)$ and $u_0 < 1/(2\lambda X)$. As λ increases, we trade spatial resolution for temporal resolution.

The experimental results that have been reported in the

Manuscript received February 14, 1983; revised June 22, 1983. This work was supported by the National Science Foundation under Grants ENG 78-25840 and ECS-8120759.

The author was with the Department of Electrical Engineering, University of Delaware, Newark, DE 19711. He is now with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

0096-3518/84/0200-0137$01.00 © 1984 IEEE