Distortion Measures for Speech Processing

ROBERT M. GRAY, FELLOW, IEEE, ANDRÉS BUZO, AUGUSTINE H. GRAY, JR., SENIOR MEMBER, IEEE, AND YASUO MATSUYAMA, MEMBER, IEEE

Abstract—**Several properties, interrelations, and interpretations are developed for various speech spectral distortion measures. The principle results are 1) the development of notions of relative strength and equivalence of the various distortion measures both in a mathematical sense corresponding to subjective equivalence and in a coding sense when used in minimum distortion or nearest neighbor speech processing systems; 2) the demonstration that the Itakura-Saito and related distortion measures possess a property similar to the triangle inequality when used in nearest neighbor systems such as quantization and cluster analysis; and 3) that the Itakura-Saito and normalized model distortion measures yield efficient computation algorithms for generalized centroids or minimum distortion points of groups or clusters of speech frames, an important computation in both classical cluster analysis techniques and in algorithms for optimal quantizer design. We also argue that the Itakura-Saito and related distortions are well-suited computationally, mathematically, and intuitively for such applications.**

I. Introduction

A DISTORTION measure is an assignment of a nonnegative number to an input/output pair of a system. The distortion between an input or original and an output or reproduction represents the cost or distortion resulting when that input is reproduced by that output. Such measures have a wide variety of applications in the design and comparison of systems.

To be useful, a distortion measure must possess to a certain degree the following properties: 1) it must be subjectively meaningful in the sense that small and large distortion correspond to good and bad subjective quality, respectively; 2) it must be tractable in the sense that it is amenable to mathematical analysis and leads to practical design techniques; 3) it must be computable in the sense that the actual distortions resulting in a real system can be efficiently computed.

The most common distortion measure is the traditional squared error or error power or error energy. This is largely because of its tractability and computability. For low bit rate speech (and image) systems, however, such a distortion measure does not appear to be subjectively meaningful. In particular, large distortion in a squared error sense does not necessarily imply poor quality. For example, a "shh" sound is essentially a white process and any typical or representative waveform will sound the same.

To overcome this problem, a number of distortion measures have been introduced which appear to be more subjectively meaningful than squared error. In effect, these have been measures of the difference between log spectra and have been implicitly used in the design of linear predictive speech systems. Gray and Markel [1] and the references cited therein present several such measures and discuss many of their properties and interrelations. While this paper builds upon Gray and Markel, we avoid the term "distance measures" used therein which, strictly speaking, requires symmetry and the triangle inequality. We instead use the term "distortion measure" with its less stringent requirements. In particular, if $d(x,y)$ is the distortion between x and y, then $d(x,y)$ must be nonnegative, and if $x=y$, then $d(x,y)=0$ (see, e.g., Berger [2, ch. 2 and 7]).

In addition to developing a number of properties, we look at the equivalences between various distortion measures and at their applications in nearest neighbor system designs such as quantization and cluster analysis systems. This includes the finding of nearest neighbors, the finding of centroids, and their application in the design of "optimal" systems. We also point out applications of these distortion measures that have been reported in other fields, far removed from speech. The major intent of this paper is to lay the theoretical groundwork for a forthcoming experimental paper and, hopefully, also for other work in related speech processing systems.

We emphasize that our focus here is on the mathematical properties of these distortion measures and not on the difficult and open problem of characterizing which distortion measures best measure the subjective quality of particular psychophysical characteristics and features of speech. We do argue, however, that the success of linear predictive coding (LPC) speech compression systems strongly suggests that the Itakura-Saito distortion measure is a subjectively meaningful distortion measure for the spectral shape of speech. This claim is further reinforced by the successful use of this distortion measure in the design of very low-rate speech compression systems (850 bits/s to 1200 bits/s) described in forthcoming papers [32], [37] and in medium rate (about 8 bits/s) waveform coders using universal tree encoding techniques from information theory [38].

II. Preliminaries

Let $f(\theta)$ represent a spectral density corresponding to a speech process, where θ is a normalized frequency ranging from $-\pi$ to π (where π corresponds to one-half of the sampling

Manuscript received August 7, 1979; revised February 20, 1980. This research was partially supported by the U.S. Air Force Office of Scientific Research, the Joint Services Electronics Program at Stanford University, Stanford, CA, the Consejo Nacional de Ciencia y Tecnologia of Mexico, and the Japan Society for the Promotion of Science, Murata Overseas Scholarship Fund.

R. M. Gray is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94395.

A. Buzo is with University of Mexico, Mexico City, Mexico.

A. H. Gray, Jr., is with Signal Technology, Inc., Santa Barbara, CA 93101.

Y. Matsuyama is with the Information Science Division, College of General Education, Ibaraki University, Mito, Japan.

0096-3518/80/0800-0367$00.75 © 1980 IEEE

frequency, or the folding frequency). The spectral density $f(\theta)$ is a nonnegative even function of θ, whose Fourier coefficients $r(n)$ define an autocorrelation sequence

$$f(\theta) = \sum_{n=-\infty}^{\infty} r(n)\, e^{-jn\theta} \tag{1}$$

$$r(n) = \int_{-\pi}^{\pi} f(\theta)\, e^{jn\theta} \frac{d\theta}{2\pi}. \tag{2}$$

For the development here it does not matter whether $f(\theta)$ represents an energy density spectrum or a power density spectrum. If $r(n)$ represents an autocorrelation defined by the expectation of lagged products of a wide-sense stationary process, then $f(\theta)$ is a power density spectrum. If $r(n)$ is a short-term autocorrelation sequence defined by a sum of lagged products, then $f(\theta)$ is an energy density spectrum.

For an ergodic wide-sense stationary process, the two are related. In particular, if a sequence $\{x(n), n = 0, 1, \cdots, N-1\}$ is used to define a short-term autocorrelation

$$r_1^{(N)}(n) = \sum_{k=0}^{N-|n|-1} x(k)\, x(k+|n|) \quad \text{for} \quad n = 0, 1, \cdots, N-1 \tag{3}$$

$$r(n) = \begin{cases} r_1^{(N)}(n) & \text{for } n < N \\ 0 & \text{for } n \geqslant N, \end{cases}$$

then $f(\theta)$ is an energy density spectrum. However, if the sequence $x(n)$ represents samples of a wide-sense stationary process, one can utilize the probabilistic autocorrelation

$$r_2(n) = E\,[x(k)\, x(k+n)]$$

where E is the expectation operator. In this case, taking $r(n) = r_2(n)$ would make $f(\theta)$ be a power density spectrum. The ergodic theorem states that with probability 1 for each n that

$$\lim_{N \to \infty} \frac{1}{N} r_1^{(N)}(n) = r_2(n).$$

Thus for large sample sizes $N^{-1} r_1^{(N)}(n)$ approximates $r_2(n)$ with high probability.

Associated with the autocorrelation sequence is a set of autocorrelation matrices. We shall denote for each N the $(N+1)$ by $(N+1)$ matrix $R_N(f)$ the matrix whose k, jth element is given by $r(k-j)$. The argument f and subscript N may be dropped when no confusion can arise, and R will be used. This is a "classical Toeplitz matrix" as in the Grenander and Szegö paper [3, p. 102]. We shall use here two fundamental properties of such matrices from Grenander and Szegö, which have also appeared in various guises in the literature on linear speech prediction.

The first property is as follows: for each positive integer n, there is associated with the spectral density $f = f(\theta)$ a *Toeplitz form*

$$T_n(a) = \int_{-\pi}^{\pi} \left| \sum_{k=0}^{n} a_k\, e^{-jk\theta} \right|^2 f(\theta) \frac{d\theta}{2\pi}$$

$$= \sum_{k=0}^{n} \sum_{l=0}^{n} a_k a_l r(k-l) = a' R_n(f)\, a \tag{4}$$

where the prime denotes transpose, and $a' = (a_0, a_1, \cdots, a_n)$. It is assumed here that a is real. It is usually convenient to define the polynomial $A(z)$ by

$$A(z) = \sum_{k=0}^{n} a_k z^{-k} \tag{5}$$

so that (4) can also be expressed as the average, over θ, of the product $|A(e^{j\theta})|^2 f(\theta)$, the energy (or power) resulting from passing a signal with an energy (or power) density spectrum f through a discrete time filter $A(z)$.

The minimum value of $T_n(a)$, for fixed n and f, subject to the constraint that $a_0 = 1$ will be denoted by $\sigma_f^2(n)$, and is given (see Szegö [4, ch. 11], Grenander and Szegö [3, ch. 2], Geronimus [5]) by

$$\sigma_f^2(n) = \det R_n(f) / \det R_{n-1}(f). \tag{6}$$

The actual form of the minimizing polynomial $A(z)$ which represents a can be expressed analytically in terms of orthogonal polynomials or found equivalently by standard techniques such as Levinson's algorithm. We shall denote here by $A_n(z)$ the nth order polynomial with $a_0 = 1$, which minimizes (4).

Of primary interest here is the fact that the minimizing polynomial $A_n(z)$ can be used with $\sigma_f(n)$ to model the spectral density $f(\theta)$ in Toeplitz form integrals. Let $G(z)$ be any polynomial of the form

$$G(z) = \sum_{k=0}^{n} g_k z^{-k}$$

representing a column vector g, whose transpose is given by $g' = \{g_0, g_1, \cdots, g_n\}$; then,

$$T_n(g) = \int_{-\pi}^{\pi} |G(e^{j\theta})|^2 f(\theta) \frac{d\theta}{2\pi}$$

$$= \int_{-\pi}^{\pi} |G(e^{j\theta})|^2 \frac{\sigma_f^2(n)}{|A_n(e^{j\theta})|^2} \frac{d\theta}{2\pi}. \tag{7}$$

This equation follows immediately from Geronimus [5, p. 12, (1.20)], which is a frequency domain representation of what has been called the correlation matching property of the model $\sigma_f(n)/A_n(z)$ [6], [7, ch. 2].

Two further properties from Grenander and Szegö [3] which have been utilized in the linear prediction literature are the notion of a one-step predictor error

$$\sigma_f^2 \triangleq \lim_{n \to \infty} \sigma_f^2(n) = \exp\left\{ \int_{-\pi}^{\pi} \ln\,[f(\theta)]\; d\theta/2\pi \right\}, \tag{8}$$

and the notion of spectral factorization. When clear from context, we drop the subscript f and abbreviate σ_f^2 by σ^2. Using a slight variation on their result by factoring $1/f(\theta)$ rather than $f(\theta)$, this gives

$$1/f(\theta) = |A(e^{j\theta})|^2 / \sigma^2 \tag{9}$$

as an autoregressive representation of $f(\theta)$, or a moving average representation of $1/f(\theta)$. The function $A(z)$ is the limiting result of $A_n(z)$

$$A(z) = \lim_{n \to \infty} A_n(z),$$

and, as in the case of $A_n(z)$, has no zeros on or outside of the unit circle $|z| = 1$. As a result, σ_f/A will be referred to as the autoregressive model for f and $1/A$ as the normalized autoregressive model. The Mth order autoregressive model for f will be $\sigma_f(M)/A_M$ and its normalized version $1/A_M$.

As a final preliminary, many of the distortion measures are most briefly expressed in terms of L_p norms on $(-\pi, \pi)$ (see, e.g., Ash [8, ch. 2]). For $p \geqslant 0$ the L_p norm of a complex-valued function g is defined by

$$||g||_p = \left\{ \int_{-\pi}^{\pi} |g(\theta)|^p \, d\theta/2\pi \right\}^{1/p}.$$

If $g(\theta)$ is continuous, then the limit as p goes to infinity of its L_p norm can be denoted by $||g||_\infty$ and is equal to its maximum magnitude. In addition,

$$||f||_p \leqslant ||f||_q \quad \text{if} \quad 0 < p \leqslant q. \tag{10}$$

III. Spectral Distortion Measures

All of the speech distortion measures considered here depend on their sampled speech waveforms only through their second-order properties—their sample autocorrelations or spectral models. These distortion measures are most easily defined in the spectral domain, though their evaluation is most often carried out without reference to that domain. These measures can be used to measure distortion between random processes, as well as between deterministic processes, by using the appropriate autocorrelation sequence.

A spectral distortion measure is a function of two spectral densities, f and $\hat{f}$ for example, which assigns a nonnegative number $d(f, \hat{f})$ to represent the distortion in using $\hat{f}$ to represent f. The most common of such measures are difference distortion measures where one uses an L_p norm on the difference $f - \hat{f}$. These are metrics or distances in the sense that they satisfy a symmetry requirement $d(f, \hat{f}) = d(\hat{f}, f)$ and a triangle inequality

$$d(f,g) \leqslant d(f,h) + d(h,g). \tag{11}$$

The spectral distortion measures considered here, however, depend upon difference in log spectra, or as a result, ratios of spectra and thus are ratio distortion measures in that they depend upon the ratios only

$$d(f, \hat{f}) = d(1, \hat{f}/f) = d(f/\hat{f}, 1). \tag{12}$$

As numerous such measures have been proposed and may someday prove useful in particular applications, it is important to know which distortion measures are genuinely different and which are "equivalent" in some sense. Intuitively, two distortion measures should be equivalent if either can be used for a specific application without significantly changing the resulting analysis or design. We consider here two precise forms of the equivalence of distortion measures. The first is identical to the usual notion of equivalence for distances or metrics; the second is a form of equivalence suitable for minimum distortion or nearest neighbor applications.

Given two distortion measures, d_1 and d_2, we say that d_1 is stronger than d_2 and write $d_1 \Rightarrow d_2$ if small d_1 implies small d_2. Mathematically speaking, for any $\epsilon > 0$ there is a $\delta = \delta(\epsilon)$ such that if $d_1 < \delta$ then $d_2 < \epsilon$. If $d_1 \Rightarrow d_2$ and $d_2 \Rightarrow d_1$, then we say that d_1 and d_2 are *equivalent* and write $d_1 \Longleftrightarrow d_2$. Intuitively, equivalence implies that while the actual numerical values for distortion may vary, they are measuring the same effect.

A distortion measure will be subjectively useful if "small" and "large" values correspond to "good" and "bad" subjective quality. Ideally subjective tests might yield thresholds defining good and bad.

The second notion of equivalence is for a particular operation that arises in quantization, cluster analysis, pattern classification, and identification—the finding of a nearest neighbor, a representation within a certain set that minimizes a distortion measure. In particular, if f is a given spectral density and $\hat{f}$ is a spectral density restricted to some finite or infinite reproduction set, but chosen so as to minimize the distortion measure $d(f, g)$ over that set, then two distortion measures d_1 and d_2 will be called *nearest neighbor equivalent* if they are minimized by the same spectral density $\hat{f}$ regardless of the reproduction set. Such equivalence can be very useful for computations, for one can use the computational form of the simplest such nearest-neighbor equivalent distortion to find a nearest neighbor.

If two distortion measures are equivalent in both senses discussed here, then we call them *completely equivalent*, and denote that by the symbol $\Longleftrightarrow$, so that $d_1 \Longleftrightarrow d_2$ indicates that the distortion measures d_1 and d_2 are completely equivalent. Note that the equivalence relations are fully consistent in that if d_1 is equivalent to both d_2 and d_3, then d_2 and d_3 are equivalent.

Two types of scaling will be used here: gain normalization and gain optimization. A gain normalized distortion measure is defined by

$$d^*(f, \hat{f}) \triangleq d(f/\sigma^2, \hat{f}/\hat{\sigma}^2) \tag{13}$$

where $\sigma^2 = \sigma_f^2$ and $\sigma^2 = \sigma_{\hat{f}}^2$ are the gains or one step prediction errors for f and $\hat{f}$ as defined by (8). Gain normalized distortion measures can be useful for separately considering the effects of normalized models and gain terms. A gain optimized distortion measure is defined by

$$d'(f, \hat{f}) \triangleq \min_{\lambda \geqslant 0} d(f, \lambda \hat{f}). \tag{14}$$

By definition,

$$d(f, \hat{f}) \geqslant d'(f, \hat{f}),$$

and, for ratio distortion measures satisfying (12), we can see that

$$d^*(f, \hat{f}) = d(f/\sigma^2, \hat{f}/\hat{\sigma}^2) \geqslant d'(f, \hat{f}).$$

Thus both d and d^* are stronger than d', or

$$d \Rightarrow d' \quad \text{and} \quad d^* \Rightarrow d'.$$

IV. Speech Spectral Distortion Measures

In this section we discuss several spectral distortion measures that have been proposed for speech applications. For many of these measures the interested reader can find a basic background discussion and reference list in Gray and Markel [1]. Our discussion here is focused on additional properties not previously pointed out, and applications of some of the distortion measures to other fields. In addition, various equivalences among the distortion measures are developed.

A. Log Spectral Deviation

One of the oldest distortion measures proposed for speech and the one which most closely resembles traditional difference distortion measures is formed by the L_p norm of the difference of the log spectra:

$$d_{\ln p}(f, \hat{f}) = \|\ln f - \ln \hat{f}\|_p = \|\ln (f/\hat{f})\|_p.$$

The most common choices for p are 1, 2, and ∞, yielding mean absolute, root mean square, and maximum deviation, respectively. From (10) it can be noted that the L_∞ norm is the strongest and the L_1 the weakest, so that

$$d_{\ln \infty} \geqslant d_{\ln 2} \geqslant d_{\ln 1}.$$

Because of its analytical tractability, the L_2 norm is the most popular (see, e.g., Gray, Gray, and Markel [9] and Viswanathan and Makhoul [10]). Note that $d_{\ln p}$ is a true pseudometric because of its symmetry and the triangle inequality known for L_p norms.

When the log spectral deviation is used with a nearest neighbor rule to select a spectral density $\hat{f}$ in some class of spectral densities such that $d_{\ln p}(f, \hat{f})$ is minimized for a fixed-sample spectral density f; then the resulting technique is sometimes referred to as analysis by synthesis.

B. Itakura–Saito Distortion

A distortion measure proposed by Itakura and Saito [11], [12], who termed it an "error matching measure," is

$$d_{IS}(f, \hat{f}) = \|(f/\hat{f}) - \ln (f/\hat{f}) - 1\|_1. \tag{15}$$

This measure arose at one point in Itakura and Saito's formulation of linear prediction as an approximate maximum likelihood estimation, and has also been used to illustrate the "spectral matching" properties of linear prediction and why poles are weighted more heavily than zeros in the fixed-sample spectral density (see, for example, Markel and Gray [7, ch. 6]).

By using the fact that

$$u - \ln (u) - 1 \geqslant 0$$

for all real u, d_{IS} can also be expressed in the form

$$d_{IS}(f, \hat{f}) = \int_{-\pi}^{\pi} (f/\hat{f}) \frac{d\theta}{2\pi} - \ln (\sigma^2/\hat{\sigma}^2) - 1 \tag{16}$$

where σ^2 and $\hat{\sigma}^2$ are the gains or one step prediction errors of f and $\hat{f}$ according to (8).

By using the series expansion

$$u = \exp (\ln u) = 1 + (\ln u) + \tfrac{1}{2} (\ln u)^2 + \cdots,$$

it has been noted [1] that d_{IS} will become approximately proportional to the mean-square log spectral distortion

$$d_{IS} \cong \tfrac{1}{2}\, d_{\ln 2}^2 \text{ for "small" distortion.} \tag{17}$$

By using Jensen's inequality [8] (or the fact that a geometric mean is less than or equal to an arithmetic mean), one can write

$$\int_{-\pi}^{\pi} (f/\hat{f}) \frac{d\theta}{2\pi} \geqslant \exp \int_{-\pi}^{\pi} \ln (f/\hat{f}) \frac{d\theta}{2\pi} = \sigma^2/\hat{\sigma}^2.$$

Using this in (16) illustrates the inequality

$$d_{IS}(f, \hat{f}) \geqslant d_{IS}(\sigma^2, \hat{\sigma}^2). \tag{18}$$

In words, for given spectral gains, constant spectra yield the smallest distortions.

The application of d_{IS} to linear prediction becomes more apparent if f is taken to be a sample-speech spectral density and $\hat{f}$ to be a model reproduction spectrum of the form

$$f(\theta) = \alpha / |A(e^{j\theta})|^2 \tag{19a}$$

where

$$A(z) = \sum_{k=0}^{M} a_k z^{-k} \quad \text{with} \quad a_0 = 1. \tag{19b}$$

Assuming $A(z)$ has all its roots inside the unit circle,

$$\hat{\sigma}^2 = \alpha.$$

From (16) to (4),

$$d_{IS}(f, \alpha/|A|^2) = \frac{1}{\alpha} T_M(\boldsymbol{a}) - \ln (\sigma^2/\alpha) - 1 \tag{20}$$

where $\boldsymbol{a}' = (1, a_1, a_2, \cdots, a_M)$. In order to choose $\boldsymbol{a}$ and α to minimize this expression, we see that $T_M(\boldsymbol{a})$ must be minimized (as in normal linear prediction) to give its minimum value $\sigma_f^2(M)$, occurring for $A(z) = A_M(z)$, and then α is chosen to minimize the result, occurring for $\alpha = T_M(\boldsymbol{a}) = \sigma_f^2(M)$. This is just the Toeplitz minimization problem of Section II.

As the use of the Itakura–Saito distortion measure for the selection of a nearest neighbor reproduction model is equivalent to the traditional linear prediction techniques which yield reasonable subjective quality, we argue that the Itakura–Saito distortion measure is a subjectively meaningful measure of speech distortion.

An alternative development of the Itakura–Saito distortion measure arises in the discrimination literature in statistics and has apparently not been previously observed in the speech literature. This other development has recently led to parallel applications of this distortion measure in nonspeech signal processing applications such as EEG analysis. Assume that we have two zero-mean Gaussian processes with spectral densities f and g, respectively, and let $p_f^N(x)$ and $p_g^N(x)$, $x = (x_0, x_1, \cdots, x_{N-1})$, x_k real, denote the corresponding Nth order probability densities. The Nth order discrimination information [or relative entropy, Kullback–Leibler number, cross entropy, or directed divergence (see, e.g., Kullback [13], Kailath [14], and Pinsker [15], Csiszár [16])] is defined by

$$I_N(f, g) = \int \cdots \int dx_0 \cdots dx_{N-1}\, p_f^N(x) \ln [p_f^N(x)/p_g^N(x)].$$

This function has extensive applications to the problem of discriminating between Gaussian processes (see Kullback [13] for a complete discussion) and is used in both detection and information theory applications. It is well known [13], [14],

[15] for Gaussian processes that

$$I_N(f,g) = \frac{1}{2} \ln \{\det \boldsymbol{R}_N(f)/\det \boldsymbol{R}_N(g)\} + \frac{1}{2} \operatorname{tr} \{\boldsymbol{R}_N(f)[\boldsymbol{R}_N(g)]^{-1}\} - \frac{N}{2}$$

where "tr" denotes the trace of a matrix, and $\boldsymbol{R}_N$ here is the $N \times N$ autocorrelation matrix associated with the spectral density in its argument (not the $N + 1$ matrix used in Section II).

From Pinsker [15, eq. 10.5.7], the normalized discrimination information has the limit

$$\begin{aligned} I(f,g) &= \lim_{N \to \infty} \frac{1}{N} I_N(f,g) \\ &= \frac{1}{2} \int_{-\pi}^{\pi} (f/g) \frac{d\theta}{2\pi} - \frac{1}{2} \ln (\sigma_f^2/\sigma_g^2) - \frac{1}{2} \\ &= \frac{1}{2} d_{IS}(f,g), \end{aligned}$$

that is, the Itakura–Saito distortion measure is exactly twice the asymptotic discrimination information under a Gaussian assumption. For this reason it is a potentially useful cost function or distortion for nearest neighbor discrimination of locally stationary time series where N is large so that $N^{-1} I_N(f,g)$ is approximately $I(f,g)$ and the observed sample spectra are approximately equal to the true spectra. These techniques have been proved experimentally useful on non-Gaussian data, such as in pattern classification of EEG waveforms (see, e.g., Gersch *et al.* [17], [18]). The connection of the Itakura–Saito distortion with discrimination information is explored in detail in [37], where it is shown that the Itakura–Saito distortion is a special case of Kullback's minimum discrimination information measure of the resemblance of a model to observed data [13, ch. 3].

C. Itakura Distortion Measure

Itakura proposed the gain-optimized Itakura–Saito distortion for use in speech recognition systems. This distortion measure is defined by

$$d_I(f,\hat{f}) \triangleq d'_{IS}(f,\hat{f}) = \min_{\lambda \geqslant 0} d_{IS}(f,\lambda \hat{f}).$$

A direct substitution in (16) with $\hat{f}$ replaced by $\lambda \hat{f}$ and $\hat{\sigma}^2$ replaced by $\lambda \hat{\sigma}^2$ yields the result

$$d_I(f,\hat{f}) = \ln \left\{ \int_{-\pi}^{\pi} \frac{f/\sigma^2}{\hat{f}/\hat{\sigma}^2} \frac{d\theta}{2\pi} \right\}.$$

d_I is related to d_{IS} through

$$d_{IS}(f,\hat{f}) = (\sigma^2/\hat{\sigma}^2) \exp [d_1(f,\hat{f})] - \ln (\sigma^2/\hat{\sigma}^2) - 1.$$

If σ/A and $\hat{\sigma}/A$ are the autoregressive models of f and $\hat{f}$ as in (9), then

$$d_I(f,\hat{f}) = \ln \left\{ \int_{-\pi}^{\pi} |A/\hat{A}|^2 \frac{d\theta}{2\pi} \right\} = \ln \{ \| A/\hat{A} \|_2^2 \}.$$

This has been referred to as a log likelihood ratio, and the bracketed term as a likelihood ratio because in the Itakura and Saito development for Gaussian sources and large sample size it approximates a likelihood ratio. This bracketed term has also been proposed as a form of distortion measure by Magill [21] in a speech quantization problem.

D. Model Distortion Measure

Itakura [19], Chaffee and Omura [20], [39], and, implicitly, Magill [21] proposed the distortion measure

$$d_m^*(f,\hat{f}) = \| 1 - \hat{A}/A \|_2^2$$

where A and $\hat{A}$ are the normalized autoregressive models of f and $\hat{f}$. By using the fact that both $A(z)$ and $\hat{A}(a)$ are analytic outside of the unit circle, this can also be expressed as

$$\begin{aligned} d_m^*(f,\hat{f}) &= 1 - 2\mathrm{Re}\,\{\hat{A}(\infty)/A(\infty)\} + \|\hat{A}/A\|_2^2 \\ &= \|\hat{A}/A\|_2^2 - 1 = e^{d_I(f,\hat{f})} - 1. \end{aligned} \tag{21}$$

From the monotonic relation between d_m^* and d_I, we see that they are completely equivalent. In addition, one can note by inspection that d_m^* is also the gain-normalized Itakura–Saito distortion measure, so that

$$d_m^* = d_{IS}^* \Longleftrightarrow d_I.$$

The distortion measure $d_m^* = d_{IS}^*$ was introduced by Itakura [19] as an approximation to d_I for small values of d_I, as can be seen from the exponential expansion applied to (21). d_m^* will always provide an upper bound for d_I.

This distortion is called a model distortion measure because it is a measure of how nearly two normalized models or filters A and $\hat{A}$ are to being inverses. For a further discussion of this point see [1] and [22].

For purposes of comparison, one can also define an unnormalized model distortion as

$$\begin{aligned} d_m(f,\hat{f}) &= \left\| 1 - \frac{\sigma/A}{\hat{\sigma}/\hat{A}} \right\|_2^2 \\ &= (\sigma^2/\hat{\sigma}^2)\, d_m^*(f,\hat{f}) + (1 - \sigma^2/\hat{\sigma}^2)^2 \\ &= (\sigma^2/\hat{\sigma}^2)\, d_m^*(f,\hat{f}) + d_m(\sigma^2,\hat{\sigma}^2). \end{aligned} \tag{22}$$

One can show a similar result for the Itakura–Saito distortion measure, that is

$$d_{IS}(f,\hat{f}) = (\sigma^2/\hat{\sigma}^2)\, d_m^*(f,\hat{f}) + d_{IS}(\sigma^2,\hat{\sigma}^2), \tag{23}$$

and use these results to demonstrate that d_{IS} and d_m are equivalent, but not in the nearest neighbor sense,

$$d_{IS} \Longleftrightarrow d_m.$$

Rather than use a gain normalization, one could use a gain optimization of d_m, which leads to the result

$$d'_m(f,\hat{f}) = 1 - 1/\|A/\hat{A}\|_2^2,$$

which can be shown to be completely equivalent through a monotonic functional relation to d_m^*

$$d'_m \Longleftrightarrow d_m^*.$$

E. Symmetrized Distortion Measures

In many of the previous examples the distortion measures were not symmetric in "origin" and "reproduction." In some cases one may wish to "symmetrize" such a distortion measure. One approach to doing this is to take a form of the mean of the measures, such as

$$d^{(q)}(f,\hat{f}) = \tfrac{1}{2}\ \{d(f,\hat{f})^q + d(f,\hat{f})^q\}^{1/q}$$

in Matsuyama *et al.* [22]. Observe that in this case $d^{(q)}(f,\hat{f}) \geq \frac{1}{2}\ d(f,\hat{f})$, and hence

$$d^{(q)}(f,\hat{f}) \Rightarrow d(f,\hat{f}).$$

Gray and Markel [1] proposed a symmetrized Itakura–Saito distortion measure, termed a COSH measure, of the form

$$d_{\cosh}(f,\hat{f}) \triangleq d_{IS}^{(1)}(f,\hat{f}),$$

which has two interpretations from the theory of random processes. In statistics, detection theory, and information theory one often uses the symmetrized directed divergence (which is simply called the divergence) between Nth order probability densities (see, e.g., Kullback [13], Kailath [14], and Pinsker [15]). Given the directed divergence $I_N(f, \hat{f})$ previously defined, the divergence is defined by

$$J_N(f,\hat{f}) = I_N(f,\hat{f}) + I_N(\hat{f},f).$$

As before, we saw that the directed divergence was asymptotically related to d_{IS}, so that similarly

$$\lim_{N\to\infty} \frac{1}{N} J_N(f,\hat{f}) = d_{\cosh}(f,\hat{f}).$$

A second interpretation is that twice $d_{\cosh}$ gives the generalized Ornstein or $\bar{\rho}$ distance between two Gaussian processes of spectral density $f/\hat{f}$ and $\hat{f}/f$ (see Gray *et al.* [23]). This is a distance measure between random processes and is a measure of how well two processes can be "fit together," that is, how much typical sequences of one process must be changed to make it look like a typical sequence of other processes.

Numerous other symmetric distortion measures can be formed by symmetrizing those in the preceeding section by gain normalizing or gain optimizing other symmetric distortion measures. Space does not permit the inclusion of all of these or of their many properties. However, the equivalence properties of some of these are included in the summary of equivalences to follow and the straightforward details of derivation may be found in [22].

F. Equivalence Summary

The various equivalences developed in the preceeding sections, along with those for some of the symmetric distortions, are summarized in Fig. 1. Recall that for a distortion measure d, d^* denotes the gain normalized distortion, d' the gain optimized distortion, and $d^{(1)}$ the symmetrized distortion. Recall also that always $d^{(1)} \Rightarrow d$, $d^* \Rightarrow d'$, and $d \Rightarrow d'$.

V. Nearest Neighbor Speech Processing

A nearest neighbor system is one wherein an input is compared with all members of an available reproduction class and the available reproduction that is "best" in the sense of minimizing the distortion between input and reproduction or being the nearest neighbor in the reproduction class to the input is produced as an output. Such systems are found in quantization systems, pattern recognition, cluster analysis, model fitting, hypothesis testing, estimation, and identification. Most of the literature on nearest neighbor systems deals with only metric or related distortion measures. In this and the following section, we show that the Itakura–Saito distortion is quite well suited for such applications, in spite of its apparent complexity and lack of symmetry.

$$\begin{tikzcd}
d_I^{(1)} & \Leftrightarrow & d'_{\cosh} & \Leftrightarrow & d_m^{(1)*} & \Rightarrow & d_I = d'_{IS} & \Leftrightarrow & d_m^* = d_{IS}^* & \Leftrightarrow & d'_m \\
d_{\cosh} & = & d_{IS}^{(1)} \arrow[u, Rightarrow] \arrow[d, Rightarrow] & \Leftrightarrow & d_m^{(1)} & \Rightarrow & d_{IS} & \Leftrightarrow \arrow[u, Rightarrow] & d_m & & \\
 & & d_{\ln 2} & & & & & & & &
\end{tikzcd}$$

Fig. 1. Distortion implications.

We now focus on the speech processing application, where f represents a sample spectrum from a frame of speech. We wish to represent f by a reproduction spectrum $\hat{f}$ chosen from a finite collection of allowable reproduction spectra, minimizing or nearly minimizing a distortion $d(f, \hat{f})$. In a cluster analysis the finite collection can be thought of as templates, reference patterns, or cluster centers. In speech compression, the finite collection is a set of allowed quantized spectral models, and only an index would be transmitted for synthesis. We again observe that we are considering here only the model, not the pitch.

One approach is a one-step system, where $\hat{f}$ is directly chosen as a nearest neighbor to f through the distortion measure $d(f, \hat{f})$. Chaffee and Omura [20], [39] named such one-step systems "rate-distortion coders" because they are used as mathematical models for ideal data compression systems in rate-distortion theory (see, e.g., Berger [2]). A two-step system arrives at a final reproduction $\hat{f}$ indirectly by first obtaining a reproduction $\tilde{f}$ in a much larger collection, which might be infinite in size. For example, $\tilde{f}$ can be restricted to the set of all possible Mth order autoregressive or all-pole models, while $\hat{f}$ represents a finite subset, due to quantization, of all possible Mth order autoregressive models.

In traditional LPC systems and some word recognition systems, $\tilde{f}$ represents an infinite precision autoregressive model found by implicitly minimizing $d_{IS}(f, \tilde{f})$. Equivalently, from (23) and (21), $\tilde{f}$ can also be obtained by first minimizing $d_m^*(f, \tilde{f})$ or $d_I(f, \tilde{f})$ to find the optimum normalized model, and then choosing the optimum gain for that model. To see this, observe that the minimum of $d_{IS}(f, \hat{\sigma}^2/|\hat{A}|^2)$ over all $\hat{\sigma}$ in a given set and all $\hat{A}$ in a given set is accomplished using (23) as

$$\min_{\hat{\sigma},\hat{A}} d_{IS}(f, \hat{\sigma}^2/|\hat{A}|^2) = \min_{\hat{\sigma}} \left\{ \frac{\sigma^2}{\hat{\sigma}^2} (\min_{\hat{A}} d_m^*(f, 1/|\hat{A}|^2)) + d_{IS}(\sigma^2, \hat{\sigma}^2) \right\}. \tag{24}$$

Given the spectrum $\tilde f$ obtained in the first step, the second step of the process is then carried out by quantizing $\tilde f$ to find $\hat f$ using some other implied distortion measure $d(\tilde f, \hat f)$. It seems reasonable to examine the use of the same distortion measure for both steps, provided we can demonstrate that the overall distortion $d(f, \hat f)$ is minimal or nearly minimal.

Let us restrict $\tilde f$ to be in the set of all Mth order autoregressive filters as described by (19). In that case, d_{IS} is minimized by $\tilde f = \sigma_f^2(M)/|A_M|^2$ and both d_m^* and d_I are minimized by $\tilde f = \tilde\sigma^2/|A_M|^2$ for any arbitrary gain $\tilde\sigma$. For this choice of $\tilde f$ we can replace f by $\tilde f$ in many integrals in which it occurs by using the correlation matching property of linear prediction, stated in the frequency domain by (7). In particular, if $\hat f$ is restricted to a subclass of the set of Mth order autoregressive filters as in (19) and

$$\tilde f = \sigma_f^2(M)/|A_M|^2,$$

then we find by direct use of the distortion measures that

$$d_{IS}(f, \hat f) = d_{IS}(f, \tilde f) + d_{IS}(\tilde f, \hat f) \tag{25}$$

with the identical relation holding for d_m^*. In these cases one has better than a triangle inequality since the relation is actually an equality! We observe that since d_{IS} is an asymptotic discrimination information, this result can be viewed as an asymptotic form of a special case of Csiszár's triangle equality for minimum discrimination information [16].

From (25) one can note that the overall distortion is minimized if one first minimizes $d_{IS}(f, \tilde f)$ over $\tilde f$ and then minimizes the remaining $d_{IS}(\tilde f, \hat f)$ over $\hat f$. This would be more obvious perhaps if one actually replaces $d_{IS}(f, \tilde f)$ by its minimum value

$$d_{IS}(f, \tilde f) = \ln\, [\sigma_f^2(M)/\sigma^2],$$

a value independent of the final model $\hat f$, a distortion due only to using a finite rather than infinite order model.

The remaining distortion $d_{IS}(\tilde f, \hat f)$ can then be expressed in the form

$$d_{IS}(\tilde f, \hat f) = T_M(\hat a)/\hat\sigma^2 - \ln\, [\sigma_f^2(M)/\hat\sigma^2] - 1 \tag{26}$$

where the Toeplitz form T_M is given by (4) with $\hat a$ being the vector whose components define $\hat A(z)$. It should be pointed out that the correlation matching property of (7) was used to allow the free interchange of f and the model $\tilde f$ in the integral defining T_M.

The actual evaluation of (26) for obtaining a nearest neighbor requires a numerical evaluation of $T_M(\hat a)$, which from (4) can be expressed in a form convenient for evaluation as

$$T_M(\hat a) = \sum_{k=0}^{M} \sum_{l=0}^{M} \hat a_k \hat a_l r(k-l)$$

$$= r(0)\hat r_a(0) + 2 \sum_{n=1}^{M} r(n)\hat r_a(n)$$

where

$$\hat r_a(n) \triangleq \sum_{k=0}^{M-n} \hat a_k \hat a_{k+n} \qquad \text{for} \qquad n = 0, 1, \cdots, M.$$

From (25), in the case considered, the two-step and one-step models are equivalent when the first step of the two-step modeling solves the normal linear prediction equations. While the illustration was for the Itakura-Saito distortion measure, the conclusion is the same for its gain normalized and gain optimized versions $d_m^* = d_{IS}^*$ and $d_I = d'_{IS}$, which are fully equivalent and thus yield the identical nearest neighbor normalized filter, that which minimizes $T_M(\hat a)$. The details are omitted for brevity.

VI. Product Codebooks

A variation of the previous nearest neighbor computations is possible when the reproduction class has a product form, being split into separate filter parameter and reproduction gain terms. Such product codebooks are useful for decreasing overall storage since R_1 bits for a gain codebook and R_2 bits for a filter codebook requires a total storage of $2^{R_1} + 2^{R_2}$ words, as opposed to an arbitrary codebook with $R_1 + R_2$ total bits for the gain and filter together which requires a storage of $2^{R_1+R_2}$ words. Optimal codebooks will not, however, have such a product form in general.

Presume that we choose $\hat f = \hat\sigma^2/|\hat A|^2$ so as to minimize (25) or (26) over all choices in a product codebook, that is, $\hat\sigma$ is selected from a gain codebook and $\hat A$ is selected from a codebook of normalized inverse filters. From (24) this can be accomplished in two steps. First choose $\hat A$ to minimize $d_m^*(f, 1/|\hat A|^2)$ or, equivalently, $d_I(f, 1/|\hat A|^2)$ or $T_M(\hat a)$, and thereby obtain the normalized model describing the spectral shape. Then, given the optimal normalized model, choose a gain $\hat\sigma$ from the gain codebook to minimize (26) or, equivalently, to complete the minimization in (24). It is important to note here that, for a fixed codebook having a product form, this two-step selection of the nearest neighbor indeed yields the minimum distortion model in the entire codebook. The given product codebook need not, however, be optimal for a fixed total rate constraint. In addition, we shall see that such a separation of gain and normalized model leads to some necessary approximations in the evaluation of centroids used in quantization and cluster analysis.

An alternative mathematical separation of the gain and normalized model can be expressed in the following way. Define

$$\alpha \triangleq T_M(\hat a),$$

and use (26) to note that

$$d_{IS}(\tilde f, \hat f) = d_{IS}(\tilde f, \hat\sigma^2/|\hat A|^2) = \alpha/\hat\sigma^2 - \ln\, [\sigma_f^2(M)/\hat\sigma^2]$$

and

$$d_{IS}(\tilde f, \alpha/|\hat A|^2) = d_I(\tilde f, 1/|\hat A|^2) = \ln\, [\alpha/\sigma_f^2(M)].$$

This can be combined with

$$d_{IS}(\alpha, \hat\sigma^2) = \alpha/\hat\sigma^2 - \ln\, [\alpha/\hat\sigma^2] - 1$$

to give the additive result

$$d_{IS}(\tilde f, \hat\sigma^2/|\hat A|^2) = d_I(\tilde f, 1/|\hat A|^2) + d_{IS}(\alpha, \hat\sigma^2). \tag{27}$$

At first glance it appears as though the two terms on the right-hand side of the equation have entirely separated the distortion into independent terms, the first arising only from $\hat{A}$ and the second from the gain $\hat{\sigma}$. It must be kept in mind, however, that the optimal gain term $\alpha = T_M(\hat{a})$ is itself a function of the elements of $\hat{A}$.

VII. Centroid Computation

An important computation in both quantizer design and several cluster analysis techniques is that of the generalized centroid of a group or cluster of terms. Consider a sequence of spectra $\{f_1, f_2, \cdots, f_L\}$, perhaps taken from a large number of speech frames where each spectra has been assigned to a group or cluster. Without loss of generality, we shall assume here that they have been subscripted so that the first n of these spectra $\{f_1, f_2, \cdots, f_n\}$ are all assigned to the same cluster. The generalized centroid (or center of gravity, or minimum distortion point) of the cluster with respect to a given distortion measure d is defined as the spectrum $\hat{f}$ within some allowed class, such as all fixed-order all-pole spectra which minimizes

$$D(\hat{f}) = \frac{1}{n} \sum_i d(f_i, \hat{f})$$

where the limits on the summation $i = 1$ to $i = n$ will be implied in this and the future summations. Thus, if all spectra in the cluster were to be modeled by a single reproduction spectrum, then the centroid would minimize the mean distortion.

The majority of cluster analysis literature (see, e.g., [24]-[28]) and quantizer literature (see, e.g., Linde *et al.* [29] and the references cited therein) deal with Euclidean and related distortion measures since centroids for such measures are well understood and readily computed. This is likely why in their work on word recognition with the Itakura distortion Levinson *et al.* [30] remark that, "... many of the classical clustering analysis techniques are rendered inapplicable. This required that extensive modifications of classical procedures be made." In particular, Levinson *et al.* replaced the centroid computation by a minimax selection of a reproduction for a cluster. Linde *et al.* [29] and Gray *et al.* [31] point out, however, that the centroid computation remains tractable via convex programming techniques for quite general distortion measures that need not be related to Euclidean distortion measures and need not be difference distortion measures. We shall show that the Itakura-Saito and related measures fit into this class and that the centroids are readily computable via standard speech processing techniques. This leads to specific design techniques for speech compression systems using the algorithm of Linde *et al.* [29]. This application is the subject of forthcoming papers, [32], [37] and preliminary results have been reported in [29] and [33]-[35].

Using the Itakura-Saito distortion measure, we obtain for $\hat{f} = \hat{\sigma}^2/|\hat{A}|^2$ that

$$D_{IS}(\hat{f}) = \frac{1}{n} \sum_i d_{IS}(f, \hat{f})$$

$$= \frac{1}{\hat{\sigma}^2} \int_{-\pi}^{\pi} \bar{f}(\theta) |\hat{A}(e^{j\theta})|^2 \frac{d\theta}{2\pi} - \frac{1}{n} \sum_i \ln(\sigma_i^2/\hat{\sigma}^2) - 1$$

where σ_i^2 is the gain or one-step prediction error for f_i, and $\bar{f}$ is the arithmetic mean of the spectra $\{f_i, i = 1, 2, \cdots, n\}$,

$$\bar{f} = \bar{f}(\theta) = \frac{1}{n} \sum_i f_i(\theta).$$

Let $\bar{\sigma}^2 = \sigma_{\bar{f}}^2$ be the gain of the mean spectrum $\bar{f}$. The above average can also be expressed in the form

$$D_{IS}(\hat{f}) = d_{IS}(\bar{f}, \hat{f}) + u \tag{28}$$

where u is a property of the cluster itself, not of $\hat{f}$, and is given by

$$u = \ln(\bar{\sigma}^2) - \frac{1}{n} \sum_i \ln(\sigma_i^2)$$

$$= \int_{-\pi}^{\pi} \ln(\bar{f}) \frac{d\theta}{d\pi} - \int_{-\pi}^{\pi} \left[\frac{1}{n} \sum_i \ln(f_i)\right] \frac{d\theta}{2\pi}. \tag{29}$$

From (28) we see that choosing $\hat{f}$ to minimize D_{IS} is identical to finding the nearest neighbor to $\bar{f}$, the spectral mean. That is, $\hat{\sigma}$ and $\hat{A}$ are the minimizing values of α and A, respectively, in (20) when f is given by $\bar{f}$. For calculations we simply replace the autocorrelation sequence by an average autocorrelation sequence, based upon an arithmetic mean of the autocorrelation sequences in the cluster.

In working with the gain optimized version of d_{IS}, the Itakura distortion, the problem is not so simple, for

$$D_I(\hat{f}) = \frac{1}{n} \sum_i d_I(f_i, \hat{f})$$

$$= \frac{1}{n} \sum_i \ln \left[\frac{1}{\sigma_i^2} \int_{-\pi}^{\pi} f_i |A|^2 \frac{d\theta}{2\pi}\right].$$

The presence of the logarithm means we must minimize a geometric mean rather than an arithmetic mean, a difficult task which might be accomplished by dynamic programming. A simpler solution is to minimize an arithmetic mean which will be an upper bound on the geometric mean, or equivalently, to use a normalized model distortion d_m^*, which is completely equivalent to d_I, overbounds d_I, and is an approximation to d_I for small values of distortion. This then yields

$$D_m^*(\hat{f}) = \frac{1}{n} \sum_i d_m^*(\bar{\bar{f}}_i, \hat{f}) = d_m^*(\bar{\bar{f}}, \hat{f})$$

where $\bar{\bar{f}}$ is now an average normalized spectrum

$$\bar{\bar{f}} = \bar{\bar{f}}(\theta) = \frac{1}{n} \sum_i f_i(\theta)/\sigma_i^2.$$

The evaluation of a centroid again becomes that of finding a nearest neighbor to an averaged spectra, giving an exact centroid for d_m^* and an upper bound and approximation for d_I. As the two measures are completely equivalent, the more tractable d_m^* should provide equally good or bad subjective quality.

Note that in the above cases the f_i can be either sample spectral densities or the corresponding Mth order autoregressive models. This permits the use of the computations in quantization or cluster analysis techniques in either a one-step or two-step nearest neighbor system.

Experimental studies of speech coder design utilizing nearest neighbor rules and the centroids found above may be found in Buzo *et al.* [32] for product codebooks (separate model and gain coders) and in Gray *et al.* [37] for combined model and gain codebooks.

VIII. Comments

We have developed several properties and interrelations of several distortion measures proposed for use in speech processing. In addition to pointing out equivalence relations, we demonstrated that the Itakura-Saito, Itakura, and normalized model distortion measures were surprisingly well suited for both the implementation and design of one-step and two-step nearest neighbor speech processing systems using standard speech processing algorithms. These results are intended to provide a mathematical foundation for future experimental research in nearest neighbor speech processing systems.

While nonsymmetric measures such as the Itakura distortion measure have been criticized [36] for use in applications considered here, it should be pointed out again that distortion measures need not be symmetric on theoretical grounds since the theory of data compression does not require it (see Berger [2]). Intuitively, a distortion need not be symmetric when there is a clear distinction in the meanings of the inputs to the measure, here an original and a reproduction.

The focus here has been on measures already implicitly in use in functional LPC systems and hence yielding reasonable subjective quality. No assumptions about the nature of speech are explicitly contained here. The interested reader is referred to Markel and Gray [7, ch. 6].

References

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

[2] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[3] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. Berkeley, CA: Univ. California Press, 1968.

[4] G. Szegö, *Orthogonal Polynomials*, 4th ed. Providence, RI: American Mathematical Society, 1975.

[5] L. Ya. Geronimus, *Orthogonal Polynomials*. New York: Consultants Bureau, Noble Offset Printers, 1961.

[6] J. D. Markel and A. H. Gray, Jr., "On autocorrelation with application to speech analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 69-79, Apr. 1973.

[7] ——, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[8] R. B. Ash, *Real Analysis and Probability*. New York: Academic, 1972.

[9] A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Comparison of optimal quantizations of speech reflection coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 9-23, Feb. 1977.

[10] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 309-321, June 1975.

[11] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Congr. Acoust.*, Tokyo, Japan, 1968, pp. C-17 to C-20.

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

[13] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[14] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Tech.*, vol. COM-15, pp. 52-60, Feb. 1967.

[15] M. Pinsker, *Information and Information Stability of Random Variables and Processes*, Ize. Akad. Nauk. SSSR, Moscow, 1960 (English translation: San Francisco, CA: Holden-Day, 1964).

[16] I. Csiszar, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, pp. 146-158, Feb. 1975.

[17] W. Gersch, F. Martinelli, J. Yonemoto, M. D. Low, and J. A. McEwan, "Automatic classification of electroencephalograms: Kullback-Leibler nearest neighbor rules," *Science*, vol. 205, pp. 193-195, 1979.

[18] ——, "Kullback-Leibler nearest neighbor rule classification of EEG's," submitted for publication.

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

[20] D. L. Chaffee, "Applications of rate distortion theory to the bandwidth compression of speech," Ph.D. dissertation, Univ. California, Los Angeles, 1975.

[21] D. T. Magill, "Adaptive speech compression for packet communication systems," in *Telecommun. Conf. Rec.*, IEEE, publ. 73 CHO 805-2, 2901-5, 1973.

[22] Y. Matsuyama, A. Buzo, and R. M. Gray, "Spectral distortion measures for speech compression," Information Systems Lab., Stanford Electronics Lab., Stanford Univ., Stanford, CA, Tech. Rep. 6504-3, Apr. 1978.

[23] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's $\bar{d}$-distance with applications to information theory," *Ann. Prob.*, vol. 3, pp. 315-328, Apr. 1975.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math., Stat., Prob.*, vol. 1, 1967, pp. 281-296.

[25] G. H. Ball and D. J. Hall, "ISODATA–An iterative method of multivariate analysis and pattern classification," in *Proc. IFIPS Congr.*, 1965.

[26] E. Forgey, "Cluster analysis of multivariate data–Efficiency versus interpretability of classifications," *Biometrics* (abstract), vol. 21, p. 768, 1965.

[27] M. R. Anderberg, *Cluster Analysis for Applications*. New York: Academic, 1973.

[28] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

[29] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, Jan. 1980.

[30] S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 134-141, Apr. 1979.

[31] R. M. Gray, J. C. Kieffer, and Y. Linde, "Locally optimal block quantizer design," *Inform. Contr.*, to appear.

[32] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," presented at Int. Conf., Acoust., Speech, Signal Processing, Apr. 1980 (an expanded version has been submitted for publication).

[33] R. M. Gray, A. Buzo, Y. Matsuyama, A. H. Gray, Jr., and J. D. Markel, "Source coding and speech compression," in *Proc. Int. Telemetering Conf.*, Los Angeles, CA, 1978, pp. 371-378.

[34] A. Buzo, A. H. Gray, Jr., and R. M. Gray, "Vector quantization of linear predictor parameters," presented at Joint Meeting Acoust. Soc. Amer. Acoust. Soc. Jap., Honolulu, HI, Dec. 1978.

[35] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "A two-step compression system with vector quantizing," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Washington, DC, 1979, pp. 52-55.

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

[37] R. M. Gray, A. H. Gray, Jr., G. Rebolledo and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination infor-

mation distortion measure," *IEEE Trans. Inform. Theory*, to be published.

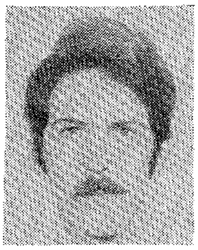
[38] Y. Matsuyama and R. M. Gray, "Universal tree encoding for speech," submitted for publication.

[39] D. L. Chaffee and J. K. Omura, "A very low rate voice compression system," Abstracts of Papers, 1974 Int. Symp. Inform. Theory, Notre Dame, IN, Oct. 1974.

Robert M. Gray (S'68-M'69-SM'77-F'80) was born in San Diego, CA, on November 1, 1943. He received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 1966, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1969.

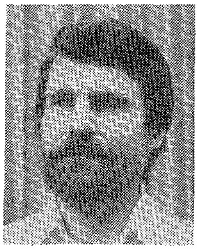
Since 1969 he has been with the Department of Electrical Engineering and the Information Systems Laboratories, Stanford University, Stanford, CA, where he is engaged in teaching and research in communication and information theory.

Dr. Gray is a member of Sigma Xi, Eta Kappa Nu, the Association for Computing Machinery, the Society for Industrial and Applied Mathematics, the Institute of Mathematical Statistics, and the Société des Ingénieurs et Scientifiques de France. He has been a member of the Board of Governors of the IEEE Professional Group on Information Theory since 1974 and an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY since September, 1977.



Andrés Buzo was born in Mexico City, Mexico, on November 30, 1949. He received the electrical and mechanical engineering degrees from the National University of Mexico, Mexico City, in 1974, and M.S. and Ph.D. degrees from Stanford University, Stanford, CA, in 1975 and 1978, respectively.

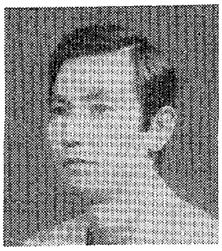
In 1978 he was with Signal Technology Inc., Santa Barbara, CA. In 1979 he joined the Instituto de Ingenieria of the National University of Mexico, where he is working on real time control systems and is engaged in research on digital signal processing.



Augustine H. Gray, Jr. (S'56-M'65-SM'77) was born on August 18, 1936. He received the S.B. and S.M. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1959, and the Ph.D. degree in engineering science from the California Institute of Technology, Pasadena, in 1964.

Since 1964 he has had numerous consulting jobs and has been a Professor with the Department of Electrical and Computer Engineering, University of California, Santa Barbara. In July 1980 he left U.C.S.B. to join Signal Technology, Inc., Santa Barbara, as a Senior Scientist, a company he has been associated with since its founding in 1977 by Dr. J. D. Markel. He coauthored (with J. D. Markel) the book *Linear Prediction of Speech* (New York: Spinger-Verlag, 1976).

Dr. Gray received the 1977 IEEE Acoustics, Speech, and Signal Processing Achievement Award for contributions to the development of linear prediction techniques.



Yasuo Matsuyama (S'77-M'78) was born in Yokohama, Japan, on March 23, 1947. He received the B.Eng., M.Eng., and Dr.Eng. degrees in electrical engineering from Waseda University, Tokyo, Japan, in 1969, 1971, and 1974, respectively.

From 1974 to 1978 he was at the Information Systems Lab, Stanford University, Stanford, CA, under the fellowships of the Japan Society for the Promotion of Science, the Murata Overseas Scholarship Foundation, and the Research Assistantship of Stanford. He received the Ph.D. degree in electrical engineering from Stanford University in 1978. He is presently a Tenured Lecturer with the Information Science Division, College of General Education, Ibaraki University, Mito, Japan.

Dr. Matsuyama is a member of the Institute of Electronics and Communication Engineers of Japan and the Society of Instrument and Control Engineers of Japan.