AT&T Bell Laboratories Technical Journal
Vol. 63, No. 8, October 1984
Printed in U.S.A.

On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation

By B.-H. JUANG*

(Manuscript received October 14, 1983)

The purpose of this paper is to discuss theoretical, as well as psychophysical, aspects of using the Itakura-Saito type of measures for evaluating the quality of coded speech. We present psychoacoustic interpretations of the measures and identify their effectiveness as well as limitations within the theoretical framework of a generalized waveform coder distortion model. The discussions then point out some specific issues to be resolved through psychoacoustic research effort.

I. INTRODUCTION

A "good" speech quality measure is central to progress in the research and development of speech processing systems. In speech coding, for example, we need a quality measure to provide insight into different distortions that are present in a coder output. If such a measure existed, it would help speech researchers identify how various kinds of distortions could be traded in order to improve the perceptual performance of the speech coder. In an engineering context, a measure that indicates the perceptual quality is a criterion to be optimized in speech coder design. Without such a measure, tuning coding schemes to achieve optimal quality is not a trivial task and the performance cannot be conveniently evaluated.

Speech quality assessment, however, involves subjective, psychological attributes of human perception, an area in which mathematicians

* AT&T Bell Laboratories.

and engineers are usually not well versed. Thus, speech quality evaluation has never been established satisfactorily in mathematical terms. The conventional signal-to-noise ratio (s/n), widely used in characterizing signal transmission/reception environments, is an ineffective measure of speech quality. Several other measurement methods and parameters, such as the isopreference method[1] and the subjective s/n,[2] have been proposed during the last two decades. General surveys of classical approaches can be found in Refs. 1 and 3. Reference 2 and its references also provide a summary of past efforts. Among these approaches, one particular class of measures based upon the Itakura-Saito measure has attracted engineers and scientists taking an analytical approach toward the problem. The Itakura-Saito measure and its variations, such as the Itakura or log likelihood ratio measure[4] and the likelihood ratio measure,[5] have been employed in noise studies by Sambur and Jayant;[6] in vocoder designs by Juang et al.[7] and Wong et al.;[8] in automatic speech recognition by Itakura[9] and Rabiner;[10] and as quality measures by Goodman et al.,[11] Crochiere et al.,[12] and Barnwell et al.[13]

Although successful applications of this class of measure are widespread in speech processing, none of them comes close to being justified as *the* speech quality measure. This paper attempts to identify the effectiveness as well as limitations of using this class of measure for speech quality within the theoretical framework of a generalized waveform coder distortion model.[14,15] We will further point out that such limitations also exist in current automatic speech recognizers that rely upon spectral matching. We then present some considerations relating to psychoacoustic studies, aiming at a better understanding of the fundamental concepts of speech quality in the presence of spectral distortion. These considerations will help direct future relevant psychoacoustic experiments for studying the dynamics of speech perception.

II. PRELIMINARIES

Let $s(i)$ and $s'(i)$ be two sampled speech signals, and let $x_n(i)$ and $x_n'(i)$ be two windowed segments, or frames, of $s(i)$ and $s'(i)$, respectively. Segments $x_n(i)$ and $x_n'(i)$ are obtained by applying a window function $w(i)$, with $w(i) = 0$ for $i < 0$ and $i \geqslant N$, to the speech signals at instance n; in particular,

$$x_n(i) = w(i)s(i + n) \qquad (1)$$

and

$$x_n'(i) = w(i)s'(i + n). \qquad (2)$$

The windowing operation greatly facilitates using spectral representations for speech analysis because speech is considered as a quasi-stationary signal. We denote the z-transform of $x_n(i)$ and $x'_n(i)$ by $X_n(z)$ and $X'_n(z)$, respectively. The Fourier transform is obtained by evaluating the z-transform on the unit circle, i.e., $z = e^{j\omega}$, and thus the notations $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$ are used to designate the Fourier transform of two windowed signals, respectively. For every such pair of spectral representations, $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$, a spectral distortion $\rho[X_n, X'_n]$ can be defined to measure the dissimilarity between $X_n(e^{j\omega})$ and $X'_n(e^{j\omega})$. In speech analysis, one particularly interesting distortion measure is the Itakura-Saito measure, which is defined as

$$\rho_{IS}[X_n, X'_n] \triangleq \int_{-\pi}^{\pi} [e^{\Lambda(\omega)} - \Lambda(\omega) - 1] \frac{d\omega}{2\pi}, \tag{3}$$

where

$$\Lambda(\omega) = \log |X_n(e^{j\omega})|^2 - \log |X'_n(e^{j\omega})|^2. \tag{4}$$

This mathematically tractable distortion measure has been successfully employed in vocoder designs.[7] Detailed analytical properties of the measure can be found in Refs. 4 and 5.

It has been shown in short-time Fourier analysis that a signal can be reconstructed from a properly time-sampled sequence of short-time Fourier transforms.[16] We can, thus, further represent the two signal sequences, $s(i)$ and $s'(i)$, by their corresponding short-time spectral sequences. Using $\oplus$ to denote the reconstruction process,

$$\{s(i)\} = .. \oplus X_{(n-1)l}(z) \oplus X_{nl}(z) \oplus X_{(n+1)l}(z) \oplus .. = \underset{n}{\oplus} X_n(z), \tag{5}$$

and

$$\{s'(i)\} = .. \oplus X'_{(n-1)l}(z) \oplus X'_{nl}(z) \oplus X'_{(n+1)l}(z) \oplus .. = \underset{n}{\oplus} X'_n(z). \tag{6}$$

In the above l is the underlying interval for short-time Fourier analysis and has been dropped in the final expressions without ambiguity. Such a representation allows us to characterize the dissimilarity between $s(i)$ and $s'(i)$ in terms of distortion measures obtained from short-time spectral representations. A distortion sequence between two speech signals is then defined as

$$\rho[s(i), s'(i)] = \{\rho_n\}, \tag{7}$$

where n is, as in (1) and (2), the frame index designating the window location, and

$$\rho_n = \rho[X_n, X'_n].$$

We will call ρ_n spectral distortion and $\{\rho_n\}$ a distortion sequence.

Extending the definition (3) to (7), then, we have a sequence of Itakura-Saito distortions.

The Itakura-Saito distortion measure defined by (3) and (4) is in fact *the* distortion measure for all-pole signal modeling; it was originally introduced as an error-matching function in maximum likelihood estimation of autoregressive spectral models.[17] Therefore, we shall confine ourselves to the analysis of Mth-order all-pole signal models despite the fact that a distortion measure could be more general. Several important results of the measure related to all-pole signal modeling are:

1. $\rho_{IS}[X_n, \sigma_n/A_n] = (\alpha_n/\sigma_n^2) + \log \sigma_n^2 - \log \alpha_{n,\infty} - 1, \qquad (8)$

where

$$\alpha_n \triangleq \int_{-\pi}^{\pi} |X_n(e^{j\omega})\cdot A_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \qquad (9)$$

$$\alpha_{n,\infty} \triangleq \exp\left\{\int_{-\pi}^{\pi} \log |X_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}\right\}, \qquad (10)$$

σ_n is a scalar, called the gain term, and

$$A_n(z) = 1 + \sum_{i=1}^{M} a_{i,n} z^{-i}. \qquad (11)$$

2. $\rho_{IS}[\sigma_n/A_n, \sigma_n'/A_n']$

$$= \frac{\sigma_n^2}{\sigma_n'^2} \int_{-\pi}^{\pi} \frac{|A_n'(e^{j\omega})|^2}{|A_n(e^{j\omega})|^2} \frac{d\omega}{2\pi} + \log \sigma_n'^2 - \log \sigma_n^2 - 1, \qquad (12)$$

which reduces to

$$\rho_{IS}[\sigma_n/A_n, \sigma_n/A_n'] = \int_{-\pi}^{\pi} \frac{|A_n'(e^{j\omega})|^2}{|A_n(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1$$
$$= \rho_{IS}[1/A_n, 1/A_n'] \qquad (13)$$

when the gain terms are identical. $A_n'(z)$ takes the same form as $A_n(z)$ in (11). In the above expressions, we have assumed that $A_n(z)$ and $A_n'(z)$ have all their roots within the unit circle. Therefore,[18]

$$\int_{-\pi}^{\pi} \log |A_n(e^{j\omega})|^2 \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \log |A_n'(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 0.$$

For clarity, we further define the likelihood ratio measure and the log likelihood ratio (or Itakura) measure as follows:

1. Likelihood ratio measure[7] $\rho_{LR}[X_n, X_n']$,

$$\rho_{LR}[X_n, X_n'] \triangleq \rho_{IS}\left[\frac{1}{A_n}, \frac{1}{A_n'}\right]; \qquad (14)$$

2. Log likelihood ratio (Itakura) measure,[4]

$$\rho_I[X_n, X_n'] \triangleq \rho_{IS}\left[\frac{\sqrt{\underline{\alpha}_n}}{\underline{A}_n}, \frac{\sqrt{\hat{\alpha}_n}}{\underline{A}_n'}\right]$$

$$= \log\left(\frac{\hat{\alpha}_n}{\underline{\alpha}_n}\right). \tag{15}$$

In defining the above two measures, $\underline{A}_n(z)$ and $\underline{A}_n'(z)$ are the *optimal* Mth-order inverse filters of $X_n(z)$ and $X_n'(z)$, respectively.[18] Furthermore,

$$\hat{\alpha}_n = \int_{-\pi}^{\pi} |X_n(e^{j\omega})\underline{A}_n'(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{16}$$

and

$$\underline{\alpha}_n = \int_{-\pi}^{\pi} |X_n(e^{j\omega})\underline{A}_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}. \tag{17}$$

Note that $\underline{\alpha}_n$ is the *minimum* Mth-order prediction residual energy pertaining to signal $X_n(z)$.

The Itakura-Saito distortion between the input and output signals of a linear system $H(e^{j\omega})$ can be easily calculated. Denoting the input power spectrum as $|X_n(e^{j\omega})|^2$, we have the output power spectrum $|X_n'(e^{j\omega})|^2 = |X(e^{j\omega})H(e^{j\omega})|^2$. Therefore,

$$\Lambda(\omega) = \log|X_n(e^{j\omega})|^2 - \log|X_n(e^{j\omega})H(e^{j\omega})|^2$$

$$= -\log|H(e^{j\omega})|^2, \tag{18}$$

and hence,

$$\rho_{IS}[X_n, X_n'] = \int_{-\pi}^{\pi}\left[\frac{1}{|H(e^{j\omega})|^2} + \log|H(e^{j\omega})|^2 - 1\right]\frac{d\omega}{2\pi}. \tag{19}$$

Of particular interest here is a class of $H(e^{j\omega})$ of the form

$$H(e^{j\omega}) = \frac{\underline{A}_n(e^{j\omega})}{B_n(e^{j\omega})}, \tag{20}$$

where $\underline{A}_n(z)$, as defined above, is the optimal Mth-order inverse filter of $X_n(z)$ and $B_n(z)$ is another Mth-order Finite Impulse Response (FIR) filter, taking the same form as (11). We also assume that $\underline{A}_n(z)$ and $B_n(z)$ both have all their roots within the unit circle. The input/output relationship of the system is illustrated in Fig. 1. Since $\underline{A}_n(z)$ is the optimal Mth-order inverse filter of $X_n(z)$, $E_n(z)$ is then the residual signal. $X_n'(z)$ is obtained by driving another all-pole filter $1/B_n(z)$ with such a residual signal. The distortion between $X_n(z)$ and

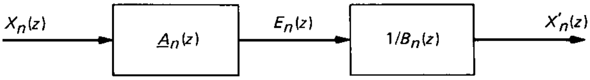


Fig. 1—A particular class of linear system in which $\underline{A}_n(z)$ is the optimal Mth-order inverse filter of $X_n(z)$.

$X'_n(z)$ under this condition is thus

$$\rho_{IS}[X_n, X'_n] = \int_{-\pi}^{\pi} \frac{|B_n(e^{j\omega})|^2}{|\underline{A}_n(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1$$

$$= \rho_{IS}\left(\frac{1}{\underline{A}_n}, \frac{1}{B_n}\right), \tag{21}$$

which is determined by the two all-pole filters, and has the same expression as the likelihood ratio measure of (14). This result gives us a convenient means of modifying a signal in order to achieve a prescribed distortion level from the original signal. Detailed discussions in Section IV are based upon this concept. It is, however, important to note that in eq. (21), $B_n(z)$ is not unique, and is not necessarily the optimal Mth-order inverse filter of the output signal $X'_n(z)$. It is simply stated that within the Mth-order autoregressive model framework, a prescribed Itakura-Saito spectral distortion can be obtained from a given signal through proper filtering operations, which will be convenient to realize.

III. A WAVEFORM CODER MODEL

Figure 2 shows a block diagram of the waveform coder distortion model used by Crochiere et al. for an interpretation of the log likelihood ratio measure.[15] This coder distortion model is composed of a time-varying linear filter $h(i)$, to model the "linearly correlated" distortions, and an additive noise source $q(i)$, to account for the nonlinear, uncorrelated distortions in the coder. Since the model attempts to split the components of distortion, it was expected that distinctively different

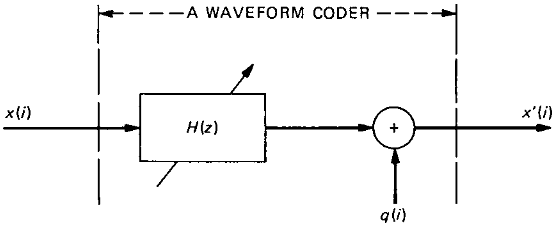


Fig. 2—Waveform coder distortion model.

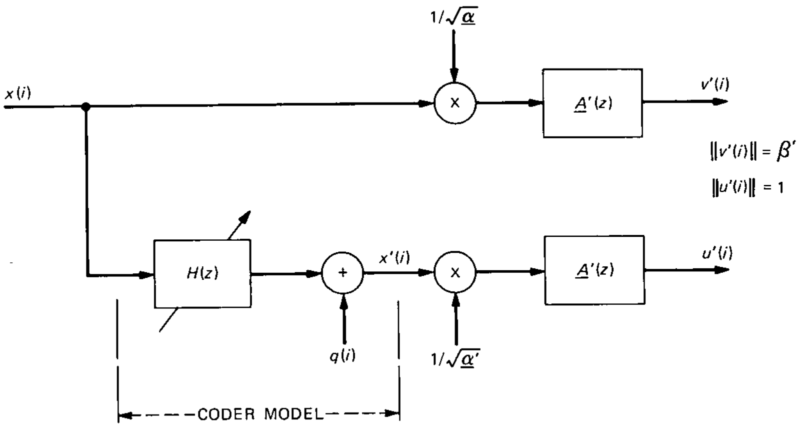


Fig. 3—Measuring coder performance with the likelihood ratio in a forward manner.

perceptual effects could be meaningfully studied separately with such a model.

Measurement of the coder performance with the likelihood ratio measure is shown in Fig. 3, which introduces the notion of inverse filtering. We use the likelihood ratio measure, rather than the Itakura-Saito measure, because we try to avoid, in the following discussions, extra complications in speech quality measurement due to amplification or attenuation. We follow the notation of Section II, except that the subscript indicating the frame index has been dropped, since, for most of the subsequent expressions, signal stationarity is assumed. We shall reinstate the frame index wherever necessary. The two parameters, α and α', are defined as in (17) by

$$\underline{\alpha} = \int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi} \tag{22}$$

and

$$\underline{\alpha}' = \int_{-\pi}^{\pi} |X'(e^{j\omega})\underline{A}'(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{23}$$

where $\underline{A}(z)$ and $\underline{A}'(z)$ are the optimal Mth-order inverse filters of $X(z)$ and $X'(z)$, respectively. In other words, $\underline{\alpha}$ and $\underline{\alpha}'$ are the minimum Mth-order prediction residual energies corresponding to $x(i)$ and $x'(i)$ sequences, respectively. The energy of $v'(i)$, denoted by β', is then

$$\beta' = \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}'(e^{j\omega})|^2 \frac{d\omega}{2\pi}. \tag{24}$$

The energy of $u'(i)$, on the other hand, is unity due to the normalization factor $1/\sqrt{\underline{\alpha}'}$ and eq. (23). The energy ratio of the two filtered signals, $v'(i)$ and $u'(i)$, is then equal to β'. By substituting (22) into (24), we have

$$\beta' = \frac{\displaystyle\int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}'(e^{j\omega})|^2 \frac{d\omega}{2\pi}}{\displaystyle\int_{-\pi}^{\pi} |X(e^{j\omega})\underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi}}. \tag{25}$$

The right-hand side of eq. (25) is the so-called likelihood ratio, and it can be reduced to

$$\beta' = \int_{-\pi}^{\pi} \frac{|\underline{A}'(e^{j\omega})|^2}{|\underline{A}(e^{j\omega})|^2} \frac{d\omega}{2\pi} \tag{26}$$

since both $\underline{A}(z)$ and $\underline{A}'(z)$ are Mth-order FIR filters, and the first $M + 1$ autocorrelation coefficients of the $\{[x(i)]/\sqrt{\underline{\alpha}}\}$ sequence are equal to those of the impulse response of $1/\underline{A}(z)$. Therefore, the likelihood ratio measure of (14) can be expressed in terms of the energy ratio of the two filtered outputs, $v'(i)$ and $u'(i)$, and

$$\begin{aligned} \rho_{LR}(X, X') \triangleq \rho_{IS}\left[\frac{1}{\underline{A}}, \frac{1}{\underline{A}'}\right] &= \int_{-\pi}^{\pi} \frac{|\underline{A}'(e^{j\omega})|^2}{|\underline{A}(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1 \\ &= \beta' - 1. \end{aligned} \tag{27}$$

Alternatively, we may replace the filter $\underline{A}'(z)$ by $\underline{A}(z)$, the inverse filter of the $x(i)$ sequence, as shown in Fig. 4. In such a case, the energy of $u(i)$, denoted by γ, is the likelihood ratio, and the distortion

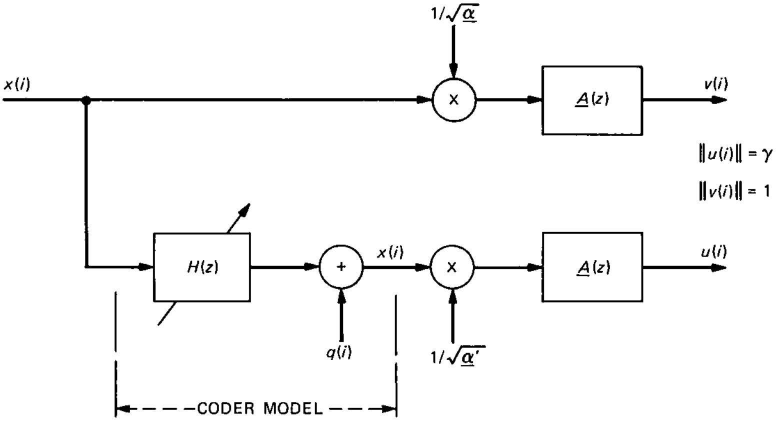


Fig. 4—Measuring coder performance with the likelihood ratio in a backward manner.

measurement is accomplished in a reversed direction, i.e.,

$$\rho_{LR}(X', X) \triangleq \rho_{IS}\left[\frac{1}{\underline{A}'}, \frac{1}{\underline{A}}\right] = \int_{-\pi}^{\pi} \frac{|\underline{A}(e^{j\omega})|^2}{|\underline{A}'(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1$$

$$= \gamma - 1. \tag{28}$$

The interpretation of the log likelihood ratio as a coder performance measure by Crochiere et al. follows the comparison order of eq. (28).[17] More specifically, the measure they discussed was log γ, instead of $\gamma - 1$. The difference between the log likelihood ratio measure and the likelihood ratio measure may be insignificant in terms of measurement. However, the likelihood ratio measure of eq. (14) appears to correspond more closely to the Itakura-Saito measure in representing the distortion relationship between the input and output signals of a particular class of linear systems. This was shown in eq. (21).

We now express the measures within the coder model. Referring to Fig. 2 and denoting the Fourier transforms of $h(i)$ and $q(i)$ by $H(e^{j\omega})$ and $Q(e^{j\omega})$, respectively, we have

$$X'(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega}) + Q(e^{j\omega}). \tag{29}$$

Furthermore, since $x(i)$ and $q(i)$ are uncorrelated,

$$|X'(e^{j\omega})|^2 = |X(e^{j\omega})H(e^{j\omega})|^2 + |Q(e^{j\omega})|^2. \tag{30}$$

For simplicity we assume that $H(z)$ does not have poles and zeros *on* the unit circle. From (24), the likelihood ratio distortion measured from $\{x(i)\}$ to $\{x'(i)\}$ is thus

$$\begin{aligned}\rho_{LR}[X, X'] &= \beta' - 1 \\ &= \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} \frac{|\underline{A}'(e^{j\omega})|^2}{|H(e^{j\omega})|^2} \{|X'(e^{j\omega})|^2 \\ &\qquad - |Q(e^{j\omega})|^2\} \frac{d\omega}{2\pi} - 1.\end{aligned} \tag{31}$$

On the other hand, the distortion measured from $\{x'(i)\}$ to $\{x(i)\}$ is

$$\begin{aligned}\rho_{LR}[X', X] &= \gamma - 1 \\ &= \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} \{|X(e^{j\omega})H(e^{j\omega})|^2 + |Q(e^{j\omega})|^2\} \\ &\qquad \cdot |\underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi} - 1.\end{aligned} \tag{32}$$

Note that $Q(e^{-j\omega})$ is the complex conjugate of $Q(e^{j\omega})$ since $q(i)$ is real.

Different distortion components of the coder may be decoupled in the following way:

1. Additive noise distortion, ρ_a, is defined when there is no correlated spectral distortion, i.e.,

$$\begin{aligned}\rho_a^{(f)} &= \rho_{LR}[X, X'] \,\big|_{|H(e^{j\omega})|=1} \\ &= \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} |\underline{A}'(e^{j\omega})|^2 \{|X'(e^{j\omega})|^2 - |Q(e^{j\omega})|^2\} \frac{d\omega}{2\pi} - 1 \\ &= \frac{\underline{\alpha}'}{\underline{\alpha}} - 1 - \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} |\underline{A}'(e^{j\omega}) Q(e^{j\omega})|^2 \frac{d\omega}{2\pi} \end{aligned} \tag{33}$$

and

$$\begin{aligned}\rho_a^{(b)} &= \rho_{LR}[X', X] \,\big|_{|H(e^{j\omega})|=1} \\ &= \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} |\underline{A}(e^{j\omega})|^2 \{|X(e^{j\omega})|^2 + |Q(e^{j\omega})|^2\} \frac{d\omega}{2\pi} - 1 \\ &= \frac{\underline{\alpha}}{\underline{\alpha}'} - 1 + \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} |\underline{A}(e^{j\omega}) Q(e^{j\omega})|^2 \frac{d\omega}{2\pi}. \end{aligned} \tag{34}$$

In the above, the superscripts, f and b, denote the forward and backward measurements, respectively.

2. Correlated spectral distortion, ρ_c, is defined when the additive noise component vanishes, i.e.,

$$\begin{aligned}\rho_c^{(f)} &\triangleq \rho_{LR}[X, X'] \,|_{Q(e^{j\omega})=0} \\ &= \frac{1}{\underline{\alpha}} \int_{-\pi}^{\pi} \frac{|X'(e^{j\omega}) \underline{A}'(e^{j\omega})|^2}{|H(e^{j\omega})|^2} \frac{d\omega}{2\pi} - 1 \end{aligned} \tag{35}$$

and

$$\begin{aligned}\rho_c^{(b)} &\triangleq \rho_{LR}[X', X] \,|_{Q(e^{j\omega})=0} \\ &= \frac{1}{\underline{\alpha}'} \int_{-\pi}^{\pi} |X(e^{j\omega}) H(e^{j\omega}) \underline{A}(e^{j\omega})|^2 \frac{d\omega}{2\pi} - 1. \end{aligned} \tag{36}$$

The above decomposition of the measure into additive noise and correlated spectral distortions provides a helpful means in cross-verification between the measure and many known perceptual attributes. In the following we shall discuss the merits as well as limitations of the above measure in measuring the perceptual quality of waveform-coded speech signals. Such discussions point to some necessary psychophysical experiments for a closer link between objective and subjective measures.

3.1 Additive noise distortion

The key contribution of the uncorrelated additive noise, $q(i)$, appears in the integral terms in (33) and (34). Let us consider (34), where the integrand involves the inverse filter $\underline{A}(z)$ for the input speech signal.

The integral

$$\int_{-\pi}^{\pi} |\underline{A}(e^{j\omega})Q(e^{j\omega})|^2 \frac{d\omega}{2\pi}$$

is minimized subject to the constraint

$$\int_{-\pi}^{\pi} |Q(e^{j\omega})|^2 \frac{d\omega}{2\pi} = P_q, \tag{37}$$

where P_q is a constant, when $\underline{A}(z)$ is the optimal (Mth-order) inverse filter of the $q(i)$ sequence. In other words, for a given noise power, the integral is minimized if *the noise has the same spectral shape as the input speech*, within the Mth-order autoregressive signal modeling framework. This appears to be in very good agreement with the results of auditory masking that has been proposed as a method for improving the perceived quality of digitally encoded speech.[19,20] The same observation can also be made on (33), where the integrand involves $\underline{A}'(z)$ instead of $\underline{A}(z)$. $\underline{A}'(z)$ is the optimal Mth-order inverse filter of the *encoded output sequence* $x'(i)$. If $q(i)$ is truly uncorrelated with $x(i)$ (recall that $|H(e^{j\omega})| = 1$ here) and has the same spectral shape as $x(i)$, then $\underline{A}'(z)$ is, in fact, identical to $\underline{A}(z)$. However, when exact shaping of noise spectra is not achievable (as in most practical coder systems), (33) and (34) lead to significantly different distortion measurements since $\underline{\alpha}'$ involves $\underline{A}'(e^{j\omega})$, which demonstrates attributes of $Q(e^{j\omega})$. The following example illustrates the difference between the forward and the backward measurements.

Consider two signals, one being tonelike and the other being white noise. These two signals are represented in terms of second-order all-pole models as $1/A_t(z)$ and $1/A_w(z)$, where

$$A_t(z) = 1 - 1.2726\, z^{-1} + 0.81\, z^{-2} \tag{38}$$

and

$$A_w(z) = 1. \tag{39}$$

The two roots of $A_t(z)$ are $0.9\, e^{\pm j\pi/4}$, which indicate a resonance at $\pi/4$ normalized frequency or at 1000 Hz when the sampling frequency is 8000 Hz. These two all-pole models have corresponding reflection coefficient vectors $\underline{k}_t$ and $\underline{k}_w$:[18]

$$\underline{k}_t^t = [k_{t1} k_{t2}] = [-0.7\ 0.81] \tag{40}$$

and

$$\underline{k}_w^t = [k_{w1} k_{w2}] = [0\ 0]. \tag{41}$$

Using eq. (7) of Ref. 7,

$$\rho_{LR}\left[\frac{1}{A_t}, \frac{1}{A_w}\right] = \frac{(k_{w1} - k_{t1})^2(1 + k_{w2})^2}{(1 - k_{t1}^2)(1 - k_{t2}^2)} + \frac{(k_{w2} - k_{t2})^2}{I - k_{t2}^2}, \tag{42}$$

we can easily calculate the distortion in each direction and obtain

$$\rho_{LR}\left[\frac{1}{A_t}, \frac{1}{A_w}\right] = 4.7 \tag{43}$$

and

$$\rho_{LR}\left[\frac{1}{A_w}, \frac{1}{A_t}\right] = 2.26. \tag{44}$$

Clearly, if measured in the forward direction, when an input tonelike signal is being distorted into white noise, the distortion is higher than vice versa. The result is reversed if the distortion is measured in the backward direction; that is, distorting an input noise signal into a tonelike signal will result in a more serious objective distortion measurement than distorting a tone-like signal into white noise. Previous studies in auditory masking demonstrated a similar asymmetry of masking between tone and noise.[21-23] In particular, it has been reported that noise masks a tone more effectively than a tone masks noise. A 1-kHz tone masked by noise that is one critical band wide typically is inaudible at a signal-to-masker ratio of −4 dB, while the corresponding ratio for noise signal masked by tone is approximately −24 dB.[24] In other words, it is easier to perceive noise in a tone than it is to perceive a tone in noise. For an objective measure to consistently predict the perceived quality, we thus would require that such a measure show higher distortion when the input tone is corrupted by noise and that it show lower distortion when input noise is distorted by an additive tone signal. Despite the slight difference between masking and distortion, forward measurements of (33) thus appear to be more justifiable. More rigorous psychoacoustic studies are obviously very important in carefully resolving this measurement direction issue.

3.2 Correlated spectral distortion

Compared to additive noise, correlated spectral distortion has not been as well studied in the past, but it is a key factor affecting the perceived quality. One well-known example is that "telephone speech", which is essentially bandlimited to the range of 200 to 3200 Hz, is considered to be of poorer quality and of lower intelligibility than the unfiltered original speech. Since correlated spectral distortion can be a result of the filtering operation, we shall discuss it using linear filtering concepts.

Linear systems can be categorized into time-invariant and time-variant systems. Accordingly, correlated spectral distortion can be time-invariant or time-variant as demonstrated in eq. (19), where the correspondence between the filtering operation and the distortion measure was established. The above-mentioned bandpass filtered speech signals, such as telephone speech, have essentially a time-invariant spectral distortion (here we are not considering tone noise, clicks, or channel variations, etc.), while Linear Predictive Coding (LPC) vocoders involve many time-variant spectral distortions, as will be discussed shortly.

The use of the Itakura-Saito type of measure for *time-invariant spectral distortions*, such as (3), (14) and (15), appears to be justifiable, at least within the short-time frame boundary where stationarity is reasonably assumed. This can be seen from the application of the likelihood ratio measure in vector quantization for voice coding.[7,8] In fact, the code words designed for vector quantization using (14) are substantially consistent with the vowel triangle of Peterson and Barney from an acoustic-phonetic point of view.[8] It also has been shown that the log likelihood ratio measure usually leads to a better recognition rate in speech recognition schemes.[10,25] (Note that the log likelihood ratio and the likelihood ratio measure make no significant difference in most speech recognition applications. The only theoretical difference is in template generation where minimization of some criterion, such as the average distortion or maximum distortion, is required.) For interests in psychoacoustic studies, however, it may be desirable to further translate the measurement into a perceptual scale that better interprets the relative perceived quality. (The complication here is the possible sound dependence on a perceptual scale. Consider the following example. Suppose X has been distorted, resulting in Y and Z. We can confidently say Y sound is closer to X sound than Z sound is, if $\rho[X, Y] < \rho[X, Z]$. However, we are not sure that Y is perceptually closer to X than Z is to W, even if $\rho[X, Y] < \rho[Z, W]$.)

Beyond the short-time stationary segment level, the time-variant distortion is a more important and complicated factor to consider in speech processing. Spectral distortion measures are defined for every pair of spectral representations. A natural extension of the distortion measure for measuring dissimilarity between time-varying signals is thus the distortion sequence is expressed by (7). Previous experiments and several reported results that help illustrate the effect of time-variant spectral distortions upon speech quality are in order.

Voice coding results in time-variant spectral distortion. The key contribution to the time variation of distortion in voice coding such as LPC is a result of parameter quantization, although the parameter analysis procedure itself may also introduce some time-variant distortion because of frame alignment, change in excitation, etc. The effect of such distortion thus can be best explained in performance comparison of different parameter quantization schemes.

The experiment in Ref. 7 that compared the distortion performance of vector and scalar quantization for LPC voice coding provides important insights in this regard. In order to conduct the so-called equal average distortion comparison in the experiment, speech signals were vocoded at a lower bit rate with vector quantization and at a higher bit rate with scalar quantization. Subjective comparison of these two sets of synthesized signals of equal average distortion showed that the vector quantization synthesis samples sounded smoother and more pleasant, and were considered of better quality. Substantial background warble was perceived in the scalar quantization samples. Differences in spectral continuity, distortion contour, and some statistics of the distortion process $\{\rho_n\}$ between the two sets of synthesis samples were then reported to explain the difference in the perceived synthesis quality. It was concluded that a coder that preserves more spectral continuity, achieves smoother distortion contour, and produces less divergent distortion statistics is better than a coder with otherwise different distortion performance, even though they yield the same average distortion. Vector quantizers appear to produce "better" distortion sequences than do scalar quantizers in LPC voice coding. The importance of considering the distortion as a *process* or *sequence* (instead of just an average distortion) and of looking into the spectral continuity (a mathematical definition of which has yet to be obtained) was thus highlighted.

The concepts of time-variant distortions and spectral continuity also raise a possible explanation for the experimental results of Tribolet et al.[26] Here, performances of four different types of waveform coders at three different bit rates were compared. An average noise-to-signal measure [eq. (2) of Ref. 26], ℓ_m, which was derived through the concepts of log likelihood ratios, was used as an objective measure to predict the subject performance. As seen from Figs. 5 and 6 (duplicated from Figs. 7 and 9a of Ref. 26), the main failures of the likelihood-ratio-derived measure are in predicting the performance of all coders, at 9.6 kb/s [in particular, Sub-band Coder (SBC) at 9.6 kb/s] and Adaptive Differential PCM (ADPCM) coder with a fixed predictor at 24 kb/s. At 9.6 kb/s all coders perform subjectively worse than objec-

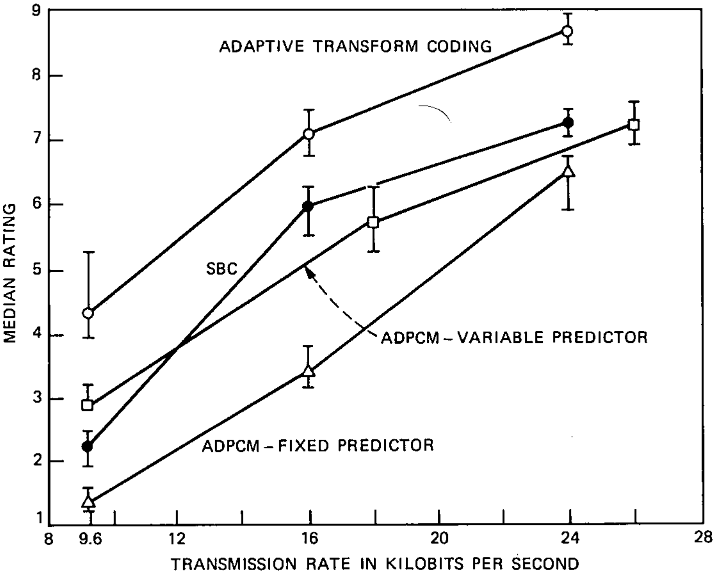


Fig. 5—Quality median rating of 12 coders (65 listeners by 4 talkers).

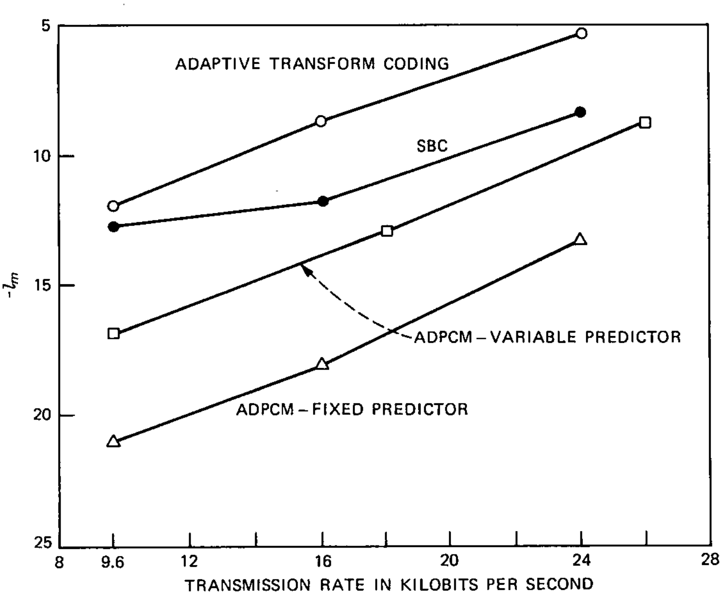
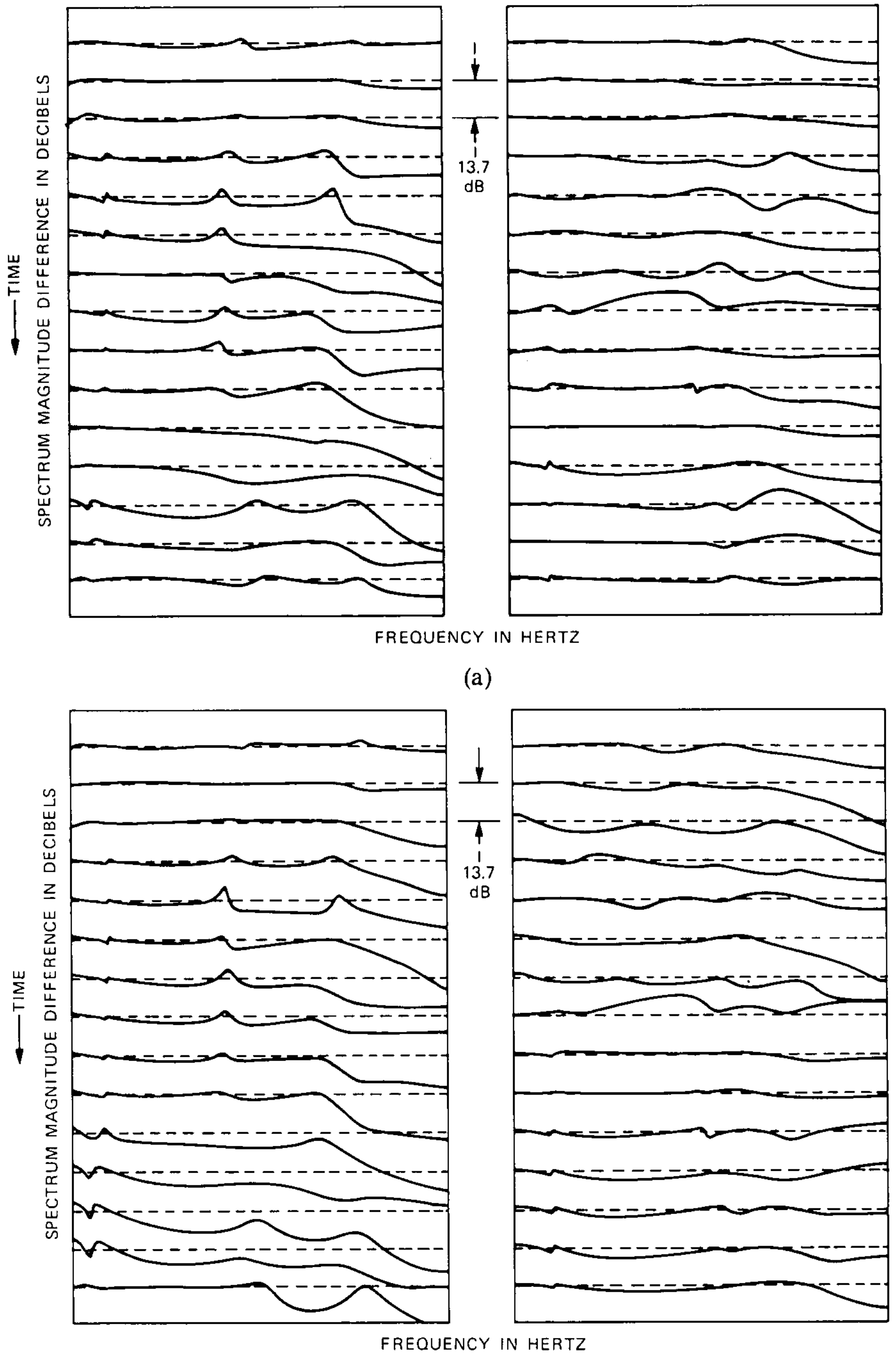


Fig. 6—Objective noise-to-signal measure, ℓ_m, averaged over 16 articulation bands for the 12 coders in Fig. 6.

tively predicted. At 9.6 kb/s, SBC is objectively very close to the Adaptive Transform Coder (ATC) but turns out to be subjectively even worse than the ADPCM with a variable predictor. At 24 kb/s, ADPCM with fixed predictor is objectively much worse than ADPCM with variable predictor, but they in fact are subjectively very close. These failures can be attributed to the fact that ℓ_m does not correctly consider the correlated spectral distortion, and more importantly, it is only an average over the entire speech sample, revealing no information on possible perceptual degradation due to time-variant spectral distortions. The outcome that all coders perform subjectively worse than objectively predicted at lower bit rates is probably a result of increased sporadic spectral distortions and reduced spectral continuity along the time axis. Sub-band coding schemes inherently preserve less spectral continuity at lower bit rate, and thus it is possible that relatively more quality degradation is perceived at 9.6 kb/s with SBC. Finally, the ADPCM coder with adaptive, variable predictor potentially introduces more spectral discontinuity, due to quantization of the predictor parameters, than does the ADPCM with a fixed, unquantized predictor.

To illustrate this, plots of the log spectral (eighth-order all-pole) difference between the original and the reconstructed speech signals are shown in Fig. 7. Coders used in Fig. 7 are ADPCM with fixed predictors and adaptive predictors, respectively. More spectral discontinuity is observed in the adaptive predictor case, particularly in the low frequency region. Therefore, even though adaptive predictors yield higher prediction gain than fixed predictors,[27] this objective advantage has been subjectively offset by the perceptual sensitivity to time-variant distortions, particularly at higher bit rates, where the effect of additive noise becomes relatively less significant. As a result, the subjective performance gap between the two coders is substantially reduced.

Similar limitations apply to automatic speech recognition schemes that use one single average or accumulative figure to represent the dissimilarity between the spectral sequences of the input speech and the stored reference template. In parallel with the concept of measuring speech quality with the segmental s/n, recognition schemes usually resort to segmentation and time warping in order to obtain better distortion or distance measurements for more accurate recognition decisions. Nevertheless, segmentation schemes produce hard segmental boundaries, instead of natural, soft transitions, and are never completely reliable. The original problem of measuring the dissimilarity between time-varying signals thus has never been entirely solved.



(a)

(b)

Fig. 7—Log spectral difference between the original and reconstructed signals: (a) with a fixed, unquantized predictor; (b) with an adaptive, quantized predictor.

The above considerations clearly point out the necessity of psychophysical experiments for developing a better speech quality measure. Specifically, with regard to using the Itakura-Saito type of measures, issues to be further studied are: the measurement direction, the feasibility of characterizing subjective quality by distortion sequences, and the incorporation of some transitive functions into the distortion measure to account for spectral continuity. In light of the analytical features of the Itakura-Saito type of measures, research on these issues appears to be vitally important to an analytical speech perception model.

IV. SOME CONSIDERATIONS IN FUTURE PSYCHOPHYSICAL STUDIES

It is beyond the scope of this paper to propose and discuss in detail the psychophysical experiment procedures necessary to answer all the questions above. It is, however, appropriate to address one of the difficulties in psychoacoustic experiment designs here. In addition, we shall propose to consider a class of transitive functions to be used in defining the spectral continuity measure.

4.1 Inverse filtering as a tool

One of the fundamental difficulties in designing psychoacoustic experiments is the control of test stimuli. How to characterize and control the test signals is obviously not a simple matter when the stimuli are real running speech signals. In parallel with this problem is the difficulty in defining a refined speech production model that, at least, adequately describes the real speech production mechanism.

In studying perceptual responses to various spectral distortions in order to better analytically and dynamically characterize speech quality with the Itakura-Saito measure, the difficulty fortunately can be greatly alleviated. In particular, the result of (21) allows us to modify a speech signal conveniently to meet the prescribed distortion requirements. Clearly, if

$$\begin{aligned}\{s'(i)\} &= \bigoplus_n X'_n(z)\\ &= \bigoplus_n \frac{E_n(z)}{B_n(z)}\\ &= \bigoplus_n X_n(z)\,\frac{A_n(z)}{B_n(z)},\end{aligned} \tag{45}$$

then

$$\rho[s(i), s'(i)] = \left\{\rho_{IS}\left[\frac{1}{A_n}, \frac{1}{B_n}\right]\right\}_n. \tag{46}$$

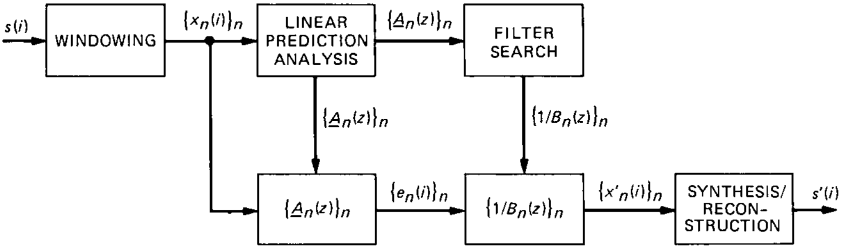


Fig. 8—Signal modification procedures to achieve prescribed distortion characteristics.

Figure 8 illustrates the modification procedure. The speech signal is first inverse filtered by $\underline{A}_n(z)$ to obtain the residual $E_n(z)$, which then drives the chosen filter $1/B_n(z)$ to form the desired signal.

Choosing $1/B_n(z)$ such that

$$\rho_{IS}\left[\frac{1}{\underline{A}_n}, \frac{1}{B_n}\right] \cong \rho,$$

a prescribed value can be made simple if we have a good-sized vector code book, as designed in vector quantization.[7] The search for $1/B_n(z)$ is then *quantum-selectively finite,* although there are theoretically infinite number of all-pole filters. Also, the test stimuli designed according to (45) are free from excitation variations, such as fundamental frequency changes, that are better considered separately.

4.2 Spectral continuity

As discussed above, spectral continuity is an important factor affecting the perceived quality of speech signals. Speech signals carry distinctive time-frequency or spectral transition patterns. Phonetic manifestation in articulated speech signals could be very fast, like /str/ in "strange", or sustainingly slow, like /i/ in "eat". To avoid complications due to such an inherent nonuniform spectral change, Ref. 7 used the model error spectral sequence $\{\Delta_n(\omega)\}$, defined by

$$\Delta_n(\omega) = \log \frac{1}{|\underline{A}_n(e^{j\omega})|^2} - \log \frac{1}{|\hat{A}_n(e^{j\omega})|^2}, \tag{47}$$

where $\hat{A}_n(z)$ is a quantized version of $\underline{A}_n(z)$, to illustrate the difference of the ability of various quantization schemes in preserving spectral continuity. The rationale was based upon the fact that the ultimate spectral continuity to be retained is the inherent spectral transition pattern, and that if a coder produces spectral distortion that is independent of time, that is,

$$\Delta_n(\omega) = \Delta(\omega) \quad \text{for all } n, \tag{48}$$

then the time-variant spectral distortion is completely eliminated. While $\{\Delta_n(\omega)\}$ adequately explained the spectral continuity differences, more rigorous alternatives are necessary for, at least, the following reasons: (1) the variation in $\Delta_n(\omega)$ along the frequency axis, ω, as well as the time axis, n, is often so substantial that it is difficult to use only (47) to define a spectral continuity measure; (2) it was never concluded that the change in $\Delta_n(\omega)$ along the time axis, if regarded as an indication of spectral smoothness, is indeed perceptually independent of the spectral transition pattern of the speech signal.

Before we can completely characterize the spectral continuity along both the frequency and time axis, we would like to propose to tentatively consider two transitive functions that indicate the spectral changes in a speech signal as a function of time. The notion of eq. (21), measuring the distortion between two all-pole spectra, is emphasized in defining such transitive functions. Denoted by $\phi_f(k)$, the forward transitive function is defined by

$$\phi_f(k) \triangleq \sum_{n=0}^{\infty} e^{-n\lambda_f} \rho_{IS} \left[\frac{1}{\underline{A}_k}, \frac{1}{\underline{A}_{k-n}} \right], \tag{49}$$

where λ_f is a time constant and, $\underline{A}_k(z)$ and $\underline{A}_{k-n}(z)$ are the optimal Mth-order inverse filters of $X_k(z)$ and $X_{k-n}(z)$, respectively. $\phi_f(k)$ measures the all-pole spectral change in the speech signal in a forward manner, i.e., it measures the distortion resulting from replacing the current spectral envelope with previous spectral envelops. Characteristic changes in excitation, such as the pitch inflection, are not actively considered in $\phi_f(k)$, although they may affect the estimation of all-pole spectral models. One interpretation of measuring the transition in speech by the distortion between all-pole models instead of speech spectra is that we try to keep the current excitation signal unchanged, as if it were present in the previous segments as implied by eq. (21). We also assume that the time constant λ_f, accounting for short-time auditory memory,[28] is independent of the particular sound that is articulated and perceived.

Similarly, we define the backward transitive function $\phi_b(k)$ as

$$\phi_b(k) \triangleq \sum_{n=0}^{\infty} e^{-n\lambda_b} \rho_{IS} \left[\frac{1}{\underline{A}_{k-n}}, \frac{1}{\underline{A}_k} \right]. \tag{50}$$

Note that if the distortion measure were symmetrical and if $\lambda_f = \lambda_b$, the two transitive functions would be identical. The appropriateness of these functions remains to be studied.

The transitive functions are to be regarded as part of the speech signal. When a speech signal is distorted because of processing or encoding, the corresponding transitive functions are distorted also. The distortion, or noise, in the transitive functions thus provides a measure of the time-variant spectral distortion that affects the spectral continuity in the original signal. Further research effort, of course, is necessary to verify the suitability of these functions or to develop a better spectral continuity measure. We feel that the concept in (49) and (50) provides a good starting point.

V. CONCLUSION

While the Itakura-Saito distortion measure and its variations have been widely employed and are considered promising in characterizing speech quality,[15] limitations in such measures still exist and have been identified within the theoretical framework of a generalized waveform coder distortion model in the above discussion. This type of measure is inherently nonsymmetric and therefore, in measuring distortions, a proper measurement direction needs to be determined. Subjective quality evaluation involves perceptual response to various degrees of distortion that has to be considered as a time function or a stochastic process. The feasibility of describing the subjective quality by finite-order statistics of the distortion process is to be studied. Furthermore, evidence shows that speech spectral continuity is also a key, if not the most important, factor affecting the subjective quality and thus, the speech spectral transition pattern should be regarded as a vital part of the speech signal. An even more fundamental and difficult task is, then, the incorporation of the spectral transition patterns into the rather static measurements of the Itakura-Saito distortion. Psychoacoustic studies are necessary to resolve these issues.

REFERENCES

1. W. A. Munson and J. E. Karlin, "Isopreference Method for Evaluating Speech-Transmission Circuits," J. Acoust. Soc. Amer., *34* (1962), pp. 762–74.
2. M. Nakatsui and P. Mermelstein, "Subjective Speech-to-Noise Ratio as a Measure of Speech Quality for Digital Waveform Coders," J. Acoust. Soc. Amer., *72*, No. 4 (1982), pp. 1136–44.
3. M. H. L. Hecker and N. Guttman, "Survey of Methods for Measuring Speech Quality," J. Aud. Eng. Soc., *15* (1976) pp. 400–3.
4. A. H. Gray, Jr. and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-24* (October 1976), pp. 380–91.
5. R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion Measures for Speech Processing," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-28* (August 1980), pp. 367–376.
6. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis From Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-24* (December 1976), pp. 448–94.
7. B.-H. Juang, D. Y. Wong, and A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-30* (April 1982), pp. 294–304.
8. D. Y. Wong, B.-H. Juang, and A. H. Gray, Jr., "An 800 Bits/s Vector Quantization LPC Vocoder," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-30* (October 1982), pp. 770–80.
9. F. Itakura, "Minimum Predication Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-23* (February 1975), pp. 67–72.
10. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Acoustics, Speech, Signal Processing, *ASSP-26* (February 1978), pp. 34–42.
11. D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman,

"Objective and Subjective Performance of Tandem Connections of Waveform Coders With an LPC Vocoder," B.S.T.J., *58,* No. 3 (March 1979), pp. 601–29.
12. R. E. Crochiere, L. R. Rabiner, N. S. Jayant, and J. M. Tribolet, "A Study of Objective Measures for Speech Waveform Coders," in Proc. 1978 Zurich Seminar on Digital Commun., pp. H1.1–7.
13. T. P. Barnwell, III, A. M. Bush, R. M. Mersereau, and R. W. Schafer, "Speech Quality Measurement," Georgia Inst. Technol., Atlanta, Tech. Rep. E21-655-77-TB-1, June 1977.
14. M. R. Aaron, J. S. Fleischman, R. W. McDonald, and E. N. Protonatarios, "Response of Delta Modulation to Gaussian Signals," B.S.T.J., *48,* No. 5 (May–June 1969), pp. 1165–95.
15. R. E. Crochiere, J. M. Tribolet, and L. R. Rabiner, "An Interpretation of the Log Likelihood Ratio as a Measure of Waveform Coder Performance," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-28* (June 1980), pp. 318–23.
16. J. B. Allen, "Short-Term Spectral Analysis and Synthesis and Modification by Discrete Fourier Transform," IEEE Trans. Acoustics, Speech, Signal Processing, *ASSP-25* (June 1977), pp. 235–8.
17. F. Itakura, "Speech Analysis and Synthesis Systems Based on Statistical Method," Doctor of Engineering Dissertation (Department of Engineering, Nagoya University, Japan, 1972). (In Japanese).
18. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* New York: Springer-Verlag, 1976.
19. M. R. Schroeder, B. A. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," J. Acoust. Soc. Amer., *66* (1979), pp. 1647–52.
20. B. J. McDermott and C. Scagliola, unpublished work.
21. J. L. Hall, unpublished work.
22. J. L. Hall and M. R. Schroeder, "Loudness of Noise in the Presence of Tones: Measurements and Non-linear Model Results," in *Psychophysical, Physiological, and Behavioral Studies in Hearing,* G. van den Brink and F. A. Bilsen, Eds., Delft, The Netherlands: Delft University Press, 1980, pp. 329–32.
23. R. P. Hellman, "Asymmetry of Masking Between Tones and Noise," Percept. Psychophys., *11* (1972), pp. 241–6.
24. J. Zwislocki, "Analysis of Some Auditory Characteristics," in *Handbook of Mathematical Psychology,* Vol. 3, R. D. Luce, R. R. Bush, and E. Galanter, Eds, New York: Wiley, pp. 1–97.
25. B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," J. Acoust. Soc. Amer., Supplement 1, *72* (Fall 1982), p. 531.
26. J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A Comparison of the Performance of Four Low-Bit-Rate Speech Waveform Coders," B.S.T.J., *58,* No. 3 (March 1979), pp. 699–712.
27. P. Noll, "A Comparative Study of Various Schemes for Speech Encoding," B.S.T.J., *54,* No. 9 (November 1975), pp. 1597–1614.
28. G. A. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits in Our Capacity for Processing Information," Psychol. Rev., *63* (1956), pp. 81–97.

AUTHOR

Biing-Hwang Juang, B.Sc. (Electrical Engineering), 1973, National Taiwan University, Republic of China; M.Sc. and Ph.D. (Electrical and Computer Engineering), University of California, Santa Barbara, 1979 and 1981, respectively; Speech Communications Research Laboratory (SCRL), 1978; Signal Technology, Inc., 1979–1982; AT&T Bell Laboratories, 1982; AT&T Information Systems Laboratories, 1983; AT&T Bell Laboratories, 1983—. Before joining AT&T Bell Laboratories, Mr. Juang worked on vocal tract modeling at Speech Communications Research Laboratory, and on speech coding and interference suppression at Signal Technology, Inc. Presently, he is a member of the Acoustics Research department, where he is researching speech communications techniques and stochastic modeling of speech signals.