A Frequency-Weighted Itakura Spectral Distortion Measure and Its Application to Speech Recognition in Noise

FRANK K. SOONG, MEMBER, IEEE, AND MAN MOHAN SONDHI

Abstract—**The performance of a recognizer based on the Itakura distortion measure deteriorates when speech signals are corrupted by noise, especially if it is not feasible to train and to test under similar noise conditions. To alleviate this problem, we consider a more noise-resistant, weighted spectral distortion measure which weights the high SNR regions in frequency more than the low SNR regions. For the weighting function we choose a "bandwidth-broadened" test spectrum; it weights spectral distortion more at the peaks than at the valleys of the spectrum. The amount of weighting is adapted according to an estimate of SNR, and becomes essentially constant (i.e., uniform weighting) in the case of clean speech. The new measure keeps the same dot product form as the original Itakura distortion measure in the autocorrelation domain. It has been tested on a 10 speaker, isolated digit database in a series of speaker independent speech recognition experiments. Additive white Gaussian noise is used to simulate different SNR conditions (from 5 dB to clean speech). The new measure performs as well as the original unweighted Itakura distortion measure at high SNR's and significantly better at medium to low SNR's. At an SNR of 5 dB, the new measure achieves a digit error rate of 12.4 percent while the original Itakura distortion gives an error rate of 27.6 percent. The equivalent SNR improvement achieved by using the proposed weighted Itakura distortion at low SNR's is about 5–7 dB.**

I. INTRODUCTION

THE short-time spectral envelope of clean speech signals can be well characterized by an autoregressive (AR) model. The modeling procedure was cast into a maximum likelihood estimation framework by Itakura and Saito [1] and used successfully in linear predictive coding (LPC) of speech. In the same paper, maximizing the likelihood function in LPC was shown to be equivalent to minimizing the Itakura–Saito distortion measure. The Itakura–Saito distortion measure between a short-time spectral periodogram and an AR model spectrum is given by

$$d_{IS} = \int_{-\pi}^{\pi} \left[\frac{|X(\omega)|^2}{|S(\omega)|^2} + \log \frac{|S(\omega)|^2}{|X(\omega)|^2} - 1 \right] \frac{d\omega}{2\pi}. \quad (1)$$

Here $|X(\omega)|^2$ is the periodogram or the short-time speech spectrum and $|S(\omega)|^2$ is the spectrum of an AR model defined by

$$|S(\omega)|^2 = \frac{\sigma^2}{|1 + a_1 e^{-j\omega} + a_2 e^{-j2\omega} + \cdots + a_p e^{-jp\omega}|^2} \quad (2)$$

where σ and a_i are the gain and ith LPC prediction coefficient of the pth-order LPC model, respectively.

Although the above distortion measure is well suited for LPC speech coding as well as for LPC front-end analysis in speech recognition, it is not very appropriate for comparing two given AR spectra in speech recognition because the measure is sensitive to the LPC gain terms [2]. In order to minimize the intrinsic gain sensitivity of the Itakura–Saito distortion measure, Itakura later developed the minimum prediction residual principle for LPC-based speech recognition [3]. The Itakura–Saito distortion between two given (reference and test) spectra is modified and the resulting new measure, commonly known as the Itakura distortion measure d_I, is given by

$$\begin{aligned} d_I &= \min_{\sigma_B > 0} d_{IS}\left(\frac{\sigma_A^2}{|A|^2}, \frac{\sigma_B^2}{|B|^2} \right) \\ &= \log \int_{-\pi}^{\pi} \frac{|1 + b_1 e^{-j\omega} + \cdots + b_p e^{-jp\omega}|^2}{|1 + a_1 e^{-j\omega} + \cdots + a_p e^{-jp\omega}|^2} \cdot \frac{d\omega}{2\pi} \end{aligned} \quad (3)$$

where $\sigma_A^2/|A|^2$ and $\sigma_B^2/|B|^2$ are the LPC spectra of two given AR models.

The measure is also known as the log likelihood ratio distortion or the gain-optimized Itakura–Saito distortion measure. It has a very simple dot product form given by

$$\begin{aligned} d_I &= \log \left(\frac{\boldsymbol{b}^T \boldsymbol{R}_A \boldsymbol{b}}{\sigma_A^2} \right) \\ &= \log \left(\sum_{i=-p}^{p} \frac{R_A(i)}{\sigma_A^2} r_B(i) \right) \end{aligned} \quad (4)$$

where $\boldsymbol{b}^T = [1, b_1, \cdots, b_p]$ is the linear prediction coefficient vector of the LPC model B, $r_B(i)$ is the ith autocorrelation coefficient of the $\boldsymbol{b}$ vector, $\boldsymbol{R}_A$ is the Toeplitz autocorrelation matrix of the model A, and $R_A(i)$ is the ith coefficient of the first row of $\boldsymbol{R}_A$.

Manuscript received January 15, 1987; revised August 10, 1987.
The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974.
IEEE Log Number 8717668.

In deriving (4) from (3) we use the so-called autocorrelation matching property [4] and the Toeplitz property of the autocorrelation matrix $\boldsymbol{R}_A$. Both the gain optimization (or the gain insensitivity) and the convenient computational form of the Itakura distortion are the two most important factors for its wide popularity in LPC-based speech recognition.

A similar spectral distortion measure called the likelihood ratio distortion measure, d_{LR}, or the gain-normalized Itakura–Saito distortion measure, is a modified Itakura distortion measure between two given LPC models where the two gain terms are normalized to 1. The measure is given by

$$d_{LR} = d_{IS}\left(\frac{1}{|A|^2}, \frac{1}{|B|^2}\right) = \int_{-\pi}^{\pi} \frac{|B|^2}{|A|^2} \frac{d\omega}{2\pi} - 1, \tag{5}$$

and except for the log function, and the constant scalar -1, is identical to d_I of (3). Since the logarithm is monotonic, the two measures are equivalent [5] except that the log likelihood ratio distortion measure has a smaller dynamic range than the likelihood ratio distortion measure. In this paper we will use the Itakura distortion measure as defined in (3).

II. A Weighted Itakura Spectral Distortion for Speech Recognition in Noise

When speech signals are corrupted by noise, the all-pole assumption is no longer valid and the signal is more appropriately modeled by an autoregressive moving average (ARMA) model [6]. For several reasons, however, AR modeling is the most commonly used procedure, even in a noisy environment. First, for speech recognition as well as for speech coding, we are interested in recovering the spectrum of the original, rather than that of the noisy signal. Second, an ARMA model estimator is intrinsically iterative and there is no convenient computational form for comparing two given ARMA models. Finally, it has been shown experimentally that if the noisy conditions are similar during training and testing, performance degrades only marginally in moderately noise conditions [7]. However, in many applications where it is not feasible to train and to test the speech recognizer under the same noisy conditions, noisy test templates must be compared to clean reference templates. In such situations, the recognition rate usually degrades dramatically and the performance is, in general, not acceptable. In order to alleviate this degradation due to mismatched training and testing conditions, we propose a new spectral distortion measure for speech recognition in noise.

Before describing the new distortion measure, let us indicate some of the disadvantages of using the Itakura distortion measure for comparing clean and noisy spectra. To illustrate the problem, we synthesized the vowel "i" (as in "bit") from an 8th-order AR model given in [8].

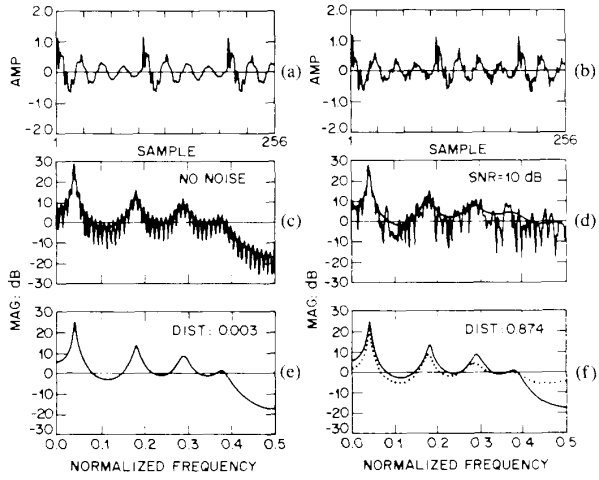

Fig. 1. A frame of clean and white noise corrupted (SNR: 10 dB) synthetic voiced speech. (a) Clean speech waveforms. (b) Noise-corrupted speech waveforms. (c) The periodogram and the LPC model spectrum of the clean speech. (d) The periodogram and the LPC model spectrum of the noisy speech. (e) The true LPC model spectrum (solid line) and the estimated LPC model spectrum (dotted line) of the clean speech. (f) The true LPC model spectrum (solid line) and the estimated LPC model spectrum (dotted line) of the noisy speech.

To minimize the effect of pitch on LPC, we chose the formant center frequencies to be multiples of the pitch frequency; and to minimize the effect of window position, we placed the window so as to start at the beginning of a pitch period. The formant center frequencies were set at 400, 1800, 2900, and 3800 Hz, and the corresponding bandwidths at 50, 140, 240, and 350 Hz. An impulse train with a period of 10 ms was used as excitation. A 12th-order LPC model was estimated using a 25.6 ms Hamming window.

The left-hand column in Fig. 1 shows various curves for the clean synthetic speech, and the right-hand column shows the same curves for the same signal with additive white Gaussian noise at a 10 dB SNR. The curves shown are: the speech waveform [(a), (b)]; the estimated LPC spectrum superimposed on the DFT periodogram [(c), (d)]; and the true AR spectrum superposed on the estimated LPC spectrum [(e), (f)].

We observe that while LPC modeling is almost perfect for this noise-free, synthesized speech frame and has an Itakura distortion measure of 0.003, the LPC spectral model of the noisy signal is quite different from the true AR spectrum and the resultant Itakura distortion is as high as 0.874. The difference between the true AR spectrum and the estimated noisy AR spectrum is larger in the spectral valleys than at the spectral peaks. While the spectrum is well approximated at the first few formant peaks, the mismatch in the valleys, especially between the 4th formant and the Nyquist frequency, makes the Itakura distance large. The spectral ratio between the true AR model spectrum and the estimated LPC spectrum clearly demonstrates this effect in Fig. 2.

The above example of synthesized speech demonstrates

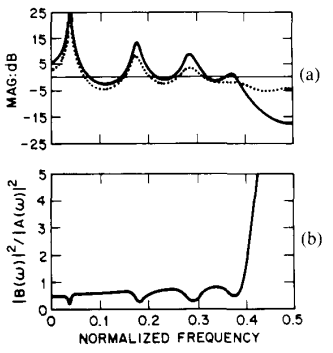

Fig. 2. (a) The true LPC model spectrum (solid line) and the estimated LPC model spectrum (dotted line) of the noisy speech in Fig. 1. (b) The spectral ratio between the two spectra in (a).

only one aspect of AR speech modeling in noise. In recognition of natural speech, we will have to compare one estimated LPC model to another estimated LPC model instead of comparing an estimated LPC model to the true model itself. Other factors, such as pitch effects in LPC modeling, the positioning of the window, and the possible inadequacy of the quasi-stationarity assumption of speech signals, further degrade the recognition performance in noise.

Since speech power is nonuniformly distributed in frequency and LPC models spectral peaks better than the spectral valleys, we propose here a weighted Itakura spectral distortion measure to exploit the relatively higher noise immunity exhibited by the spectral peaks. The spectral distortion is given by

$$d_{WI} = \log \int_{-\pi}^{\pi} F(\omega) \frac{|B(\omega)|^2}{|A(\omega)|^2} \frac{d\omega}{2\pi} \tag{6}$$

where $F(\omega)$ is a weighting function.

Let B be the reference LPC spectrum and A the test. Then one reasonable choice for $F(\omega)$ is a "bandwidth-broadened" test spectrum given by

$$\frac{1}{|A_\alpha(\omega)|^2} = \frac{1}{|1 + \alpha a_1 e^{-j\omega} + \cdots + \alpha^p a_p e^{-jp\omega}|^2} \tag{7}$$

where α, with a value between 0 and 1, is the so-called "bandwidth-broadening factor." Such bandwidth broadening has been used in speech coding for solving the bandwidth underestimation problem in LPC and for "perceptual" shaping of the quantization noise in speech coders [9], [10]. Since center frequencies of speech formants are of more perceptual importance than bandwidths and amplitudes, the proposed measure is intuitively appealing because it changes only the bandwidths of the corresponding roots (poles) of an LPC polynomial while it leaves the center frequencies intact.

With the weighting function of (7) substituted for $F(\omega)$ in (6), it is seen that d_{WI} is just the Itakura distance between $B(\omega)$ and $A_\alpha(\omega)\,A(\omega)$. It can therefore be computed in the dot product form given in (4) (with the matrix R_A suitably redefined). However, since $B(\omega)$ is of order p and $A_\alpha(\omega)\,A(\omega)$ is of order $2p$ (except in the trivial case of $\alpha = 0$), only the $(p+1) \times (p+1)$ leading minor of the matrix need be retained. The complete computation is accomplished by the following procedure.

PROCEDURE WEIGHT_AC
For the whole test utterance from frame $i = 1$ to $i = n$
DO
1) Obtain a pth-order LPC prediction error filter, $[1, a_1, \cdots, a_p]$, for each frame
2) Compute the weighted LPC filter, $[1, \alpha a_1, \cdots, \alpha^p a_p]$, for a given bandwidth broadening factor, α
3) Compute the product of the two filters to obtain a $2p$th-order new filter $[1, c_1, c_2, \cdots c_{2p}]$
4) Compute the impulse response of the $2p$th-order AR filter computed in 3, i.e., $1/C(z)$, and the corresponding pth-order autocorrelation coefficients of the impulse response
5) Solve the Toeplitz normal equations using the autocorrelation coefficients $[\bar{r}(0), \cdots, \bar{r}(p)]$ to obtain the LPC gain, $\bar{\sigma}_A^2$, and normalize each autocorrelation coefficient, $\bar{r}(i)$, by $\bar{\sigma}_A^2$

END

In the recognition phase, the computation is identical to the original unweighted Itakura distortion measure. The distortion is computed as

$$d_{WI} = \log \left(\frac{\boldsymbol{b}^T \bar{\boldsymbol{R}}_A \boldsymbol{b}}{\bar{\sigma}_A^2} \right). \tag{8}$$

The only difference between this and the d_I given by (4) is that the autocorrelation matrix, $\boldsymbol{R}_A$, and the LPC gain term, σ_A^2, are replaced by their weighted counterparts, $\bar{\boldsymbol{R}}_A$ and $\bar{\sigma}_A^2$, respectively.

Since the major computational load in speech recognition is in the computation of distortion, the proposed measure has exactly the same computational complexity as the original Itakura distortion measure except for the computation of the weighted autocorrelation coefficients. The computation of the weighted autocorrelation as given in the procedure WEIGHT_AC has to be done only once for each frame of the test signal. Therefore, overall, there is only a marginal increase in the computational requirements as compared to the unweighted Itakura distortion measure.

Intuitively, the bandwidth-broadening factor should be chosen in such a way that for the noise-free case no weighting (i.e., $\alpha = 0$) is used, and when SNR is low a larger α weighting is used. The rationale for adapting α to SNR is as follows. Any nonzero value of α produces a weighting which is larger at the peaks and smaller in the valleys. Since the Itakura distortion measure is one of the

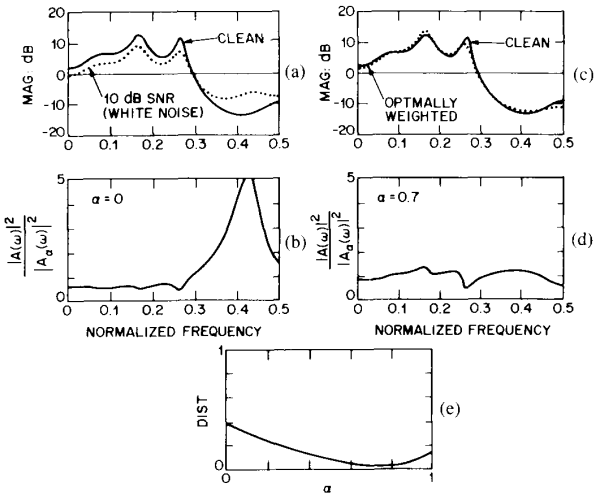

Fig. 3. (a) The LPC spectra of a frame without noise (solid line) and with noise (10 dB SNR, dotted line). (b) The spectral ratio between the two spectra in (a) as a function of frequency. (c) The LPC spectrum of (a) (solid line) and the peak weighted noise-corrupted LPC spectrum of (b) (dotted line) with an optimal weighting. (d) The spectral ratio between the two spectra in (c) as a function of frequency. (e) The weighted Itakura distortion measure as a function of α.

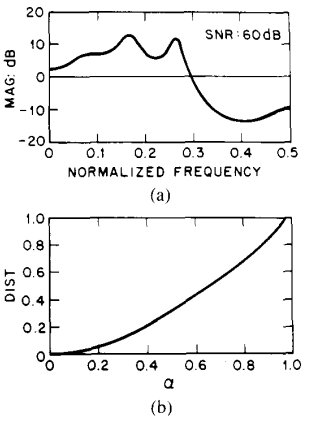

Fig. 4. (a) The LPC spectrum of the same frame in Fig. 3(a). (b) The weighted Itakura distortion as a function of α.

best known distortion measures for recognizing clean speech, it is unlikely that such a weighting will improve recognition performance. In the case of a noisy test utterance, however, a nonzero α improves the match to the reference by giving less weight to the valleys which are most affected by the noise. One would expect, of course, that there is an optimum value of α for a given SNR beyond which the match will become worse again. The experiments described in Section III show that this is indeed the case. Here we demonstrate the effect of α directly on a frame of speech. Real speech instead of synthetic speech is used for this demonstration. In Fig. 3(a) clean (solid line) and noisy (10 dB SNR, dotted line) LPC spectra are depicted. The spectral ratio between the two given spectra is shown in Fig. 3(b) where the large value of the ratio near the Nyquist frequency is, as in Fig. 2(b), due to the spectral valley mismatch. When an optimal weighting factor, $\alpha = 0.7$, is used, the weighted noisy spectrum follows the noise-free spectrum fairly well, as shown in 3(c), and the spectral ratio is more or less uniformly distributed along the frequency axis as shown in 3(d). The weighted Itakura distortion as a function of α is plotted in Fig. 3(e). A similar distortion curve is depicted in Fig. 4 for a high SNR (60 dB). The distortion curve clearly demonstrates that for a high SNR, the unweighted Itakura distortion, $\alpha = 0$, is the best choice. Any nonzero α increases the distortion and can therefore degrade the noise-free recognition performance.

In order to determine the optimum value of α as a function of the SNR, we used 500 frames of speech randomly chosen from an isolated digit database. Machine-generated zero-mean, additive white Gaussian noise samples were added to clean speech signals at seven different SNR's (−5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and ∞ dB). For each frame, the optimal α is the value that minimizes the weighted Itakura distortion between the clean AR model and the noise-corrupted AR estimate. The resultant optimal α versus SNR is depicted in Fig. 5 and tabulated in Table I. This information is used in the recognition experiments in the next section to adaptively choose the optimal α based on a frame-by-frame estimate of segmental SNR.

Two observations can be made about the statistics shown in Table I. First, at high SNR, α should be set to zero and for very low SNR (≤ 0 dB) α should be set equal to 1. The second observation is that the standard deviation is very small at these extreme SNR values, but is large at values of SNR in the range of 10 or 20 dB. Also, as shown in Fig. 3(e), the minimum of the weighted distortion as a function of α is very shallow at these SNR's. Hence, the estimate of SNR (and the corresponding α) need not be very precise. The experiments described in the next section bear this out.

III. The LPC Analysis, Database, and Experimental Results

In this section we summarize our LPC analysis front-end and the database used in the recognition experiments. We also give detailed descriptions of our experiments and results.

A. Experimental Setups and Database

The LPC front-end analysis procedure is the one usually employed in our lab [11]. The analog speech signals are first bandpass filtered and then digitized at a sampling rate of 6.67 kHz. The digitized speech is then preemphasized using a first-order filter with a transfer function of $1 - 0.95z^{-1}$ and analyzed with an 8th-order LPC analysis. A 45 ms Hamming window is used in computing the first nine autocorrelation coefficients every 15 ms. The LPC front-end and the speaker-independent, isolated word recognizer are shown at the top and the bottom of Fig. 6, respectively. The end-pointed reference or test templates

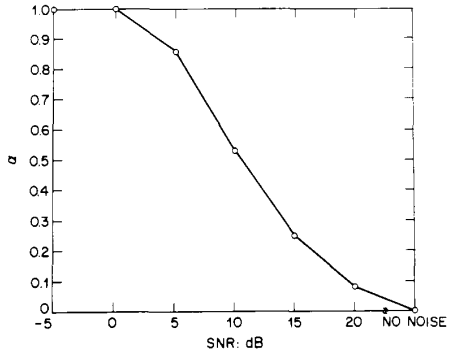

Fig. 5. The optimal α's for different segmental SNR's (from -5 dB SNR to no noise).

TABLE I
STATISTICS OF THE OPTIMAL BANDWIDTH-BROADENING FACTOR

SNR_{seg} (dB)	Optimal α	
	Mean	Standard Deviation
-5	1	0
0	0.999	0.009
5	0.86	0.11
10	0.53	0.21
15	0.25	0.15
20	0.08	0.07
Clean	0	0

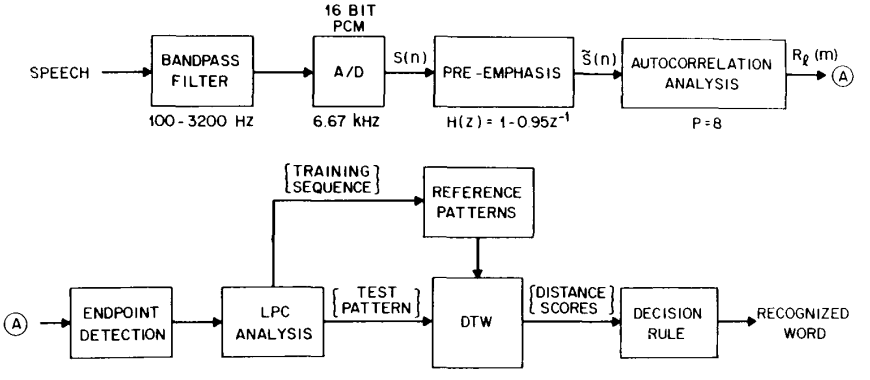

Fig. 6. The LPC front-end analysis and the DTW-based word recognizer.

are generated using the standard autocorrelation method. In the recognition phase, test templates are compared to reference templates through a standard dynamic time warping (DTW) procedure. A K nearest neighbor (KNN) decision rule chooses the word whose average DTW score of the K best word patterns is the smallest. We used KNN $= 2$ in all our experiments.

The test database used in our recognition experiments is an isolated digit database which consists of 10 different speakers (5 male and 5 female). The database was recorded over local dialed-up telephone lines in a sound booth. Each digit was repeated 10 times by each speaker. Reference templates used in our digit recognition experiments are generated from a separate training database of 100 different speakers. Twelve templates per digit are generated through the modified k-means clustering procedure of Wilpon and Rabiner [12].

For storage efficiency, the whole database was stored in an autocorrelation format instead of the original digitized samples. As a result, different noise conditions had to be simulated by adding the autocorrelations of a zero-mean, white Gaussian noise sequence to the autocorrelations of test utterances. For uncorrelated speech and noise samples, the expected value of autocorrelations of noise-corrupted speech equals the sum of the expected values of both noise and clean speech autocorrelations. However, for finite data, there could be some discrepancy between the sum of autocorrelations computed separately from noise and clean speech samples and autocorrelations computed directly from noisy samples. The effect of this approximation on the absolute recognition accuracies quoted in the next section can be evaluated only by conducting an experiment on a database of digitized speech signals. However, we believe that our estimate of the improvement in recognition accuracy, due to the use of our frequency-weighting function, is reliable.

The recognition experiments were performed at different global SNR's. The global SNR of an n-frame test utterance is defined as

$$\text{SNR} = 10 \log_{10} \frac{\frac{1}{n} \sum_{i=1}^{n} P_s(i)}{P_N} \tag{9}$$

where $P_s(i)$ is the speech signal power of the ith frame, and P_N is the noise power which is constant throughout the whole test utterance. Note that at a specified global SNR, the segmental SNR within an utterance varies from frame to frame. We believe that this experimental setup for the noisy conditions is more realistic than a constant segmental SNR arrangement.

B. Recognition Results with a Fixed Weighting

The first series of recognition experiments is used to evaluate the weighted Itakura distortion with a fixed bandwidth-broadening factor throughout the whole test utterance. The results are depicted in Fig. 7 and tabulated in Table II for different SNR's and α's. Several observations can be made about the results as follows.

1) As expected, the original unweighted Itakura distortion measure works very well at a high SNR or in a noise-free condition, but the recognition performance degrades very rapidly with decreasing SNR's. At an SNR of 5 dB, the digit error rate is 27.6 percent.

2) Using a fixed nonzero α consistently degrades the recognition in the absence of noise. The larger the spectral peak weighting (i.e., α closer to 1), the worse the recognition accuracy. An intuitive reason for this was given earlier.

3) At a low SNR of 5 dB, the recognition performance is significantly improved when a full peak weighting (i.e., $\alpha = 1$) is used. The digit error rate is reduced from 27.6 to 12.4 percent when we use $\alpha = 1$ to replace the original unweighted Itakura distortion.

4) Compared to the wide spread in digit error rates with the unweighted Itakura distortion, as shown in Fig. 7, the recognition performance curves of the weighted Itakura distortion measure bundle together more tightly. The digit recognition performance of the new weighted distortion is therefore more robust under different SNR's than that of its unweighted counterpart. Unfortunately, nonzero values of α degrade the performance at high SNR's. As discussed in the previous section, this degradation can be avoided by adjusting α according to an estimate of segmental SNR. In the next subsection we will present the recognition results obtained by using such an adaptive weighting.

C. Recognition Results with an Adaptive, SNR Dependent Weighting

In our experiments, since the noiseless data are available and the machine-generated zero-mean, white Gaussian noise is artificially added, we have full knowledge of the exact segmental SNR's for every frame in a test utterance and there is no need to estimate it from the noisy speech signal. In the following experiments we will use the known segmental SNR's (quantized to the nearest 5 dB bracket) to pick the optimal value of α from the curve in the Fig. 5. In practice, since the exact segmental SNR is not known *a priori*, it has to be estimated from the measured signals. Under the assumption that ambient noise is much more stationary than speech signals, the noise samples obtained between speech utterances can be used to estimate the background noise level. The noise characteristics are assumed to be unchanged throughout the utterance; hence, a measurement of the power level of the noisy speech signal gives an accurate estimate of the segmental SNR for each frame. The segmental SNR thus estimated is usually fairly accurate as long as the noise is reasonably stationary and end-point estimates are made conservatively (i.e., only very low level speech signals are allowed to affect the estimation of noise level). However, in order to test the resistance of the recognizer to the impreciseness of the segmental SNR estimates, in our second experiment we quantize the segmental SNR into four crude brackets and use the bracketed SNR for selecting the adaptive weighting coefficients. The four brackets of SNR's used are high (>20 dB), medium (between 10 and 20 dB), low (between 5 and 10 dB), and very low (<5 dB). Four bandwidth-broadening factors, 0, 0.25, 0.7, and 1.0 are assigned to these high, medium, low, and very low segmental SNR brackets in our experiments.

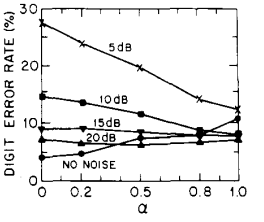

Fig. 7. The digit recognition performance using fixed α's.

TABLE II
DIGIT ERROR RATES (PERCENT) AS A FUNCTION OF FIXED WEIGHTING (I.E., α IS CONSTANT THROUGHOUT THE WHOLE TEST UTTERANCE) UNDER DIFFERENT SNR CONDITIONS
Digit Error Rate (PERCENT)

	SNR (dB)				
α	5	10	15	20	Clean
0	27.6	14.6	9	7.2	4.1
0.2	24	13.5	9	6.4	4.6
0.5	19.6	11.6	8.4	6.2	7.4
0.8	14	8.7	7.8	6.6	7.8
1	12.4	8	7.7	7.1	10.7

The results of the first experiments where the exact segmental SNR (quantized to the nearest 5 dB level) is used to choose the optimal α are given in Fig. 8 and tabulated in Table III. Also shown in Fig. 8 is the performance curve of the unweighted Itakura distortion measure. At high SNR's (>20 dB), the two curves essentially merge into one, while at low to medium SNR's, the weighted Itakura distortion measure outperforms the unweighted Itakura distortion by a fairly wide margin. The advantage in recognition performance is equivalent to a 5–7 dB SNR improvement. The result of the second experiment, where the segmental SNR measurements were quantized to four crude brackets as high, medium, low, and very low, are given in Table IV. We note that except for a slight statistical variation, the performance of the recognizer is fairly insensitive to the choice of α and the results are comparable to the results obtained when the optimal weighting is used. Consequently, we conclude that a precise segmental SNR estimate is not crucial to the robust performance of the adaptively weighted Itakura distortion measure.

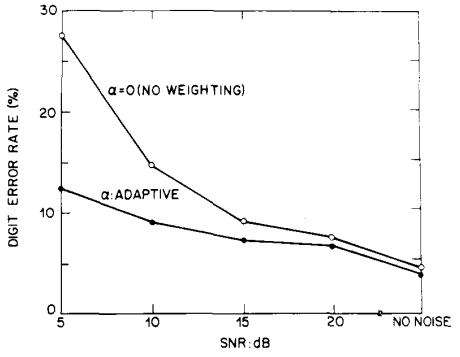

Fig. 8. The digit recognition performance using a fixed α ($\alpha = 0$) and an optimal, adaptive weighting.

TABLE III
DIGIT ERROR RATES (PERCENT) FOR AN OPTIMAL ADAPTIVE WEIGHTING UNDER DIFFERENT SNR CONDITIONS

SNR (dB)	5	10	15	20	Clean
Digit error rate (percent)	12.4	9	7.1	6.4	3.4

TABLE IV
DIGIT ERROR RATES (PERCENT) FOR A BRACKETED, ADAPTIVE WEIGHTING UNDER DIFFERENT SNR CONDITIONS

SNR (dB)	5	10	15	20	Clean
Digit error rate (percent)	12.3	9.2	7.4	6.9	3.8

IV. DISCUSSIONS

A frequency-weighted Itakura–Saito distortion measure was first proposed by Chu and Messerschmitt [13] for LPC vocoding. They used both a fixed one-pole weighting and a multiple-pole weighting in their spectral analysis of speech signals. For clean speech, they found no distinct advantage in using the weighted Itakura–Saito distortion over the unweighted one.

In speech recognition, a peak-weighted LPC distortion measure, called the weighted likelihood measure, was first proposed and investigated by Sugiyama and Shikano [14]. In their weighted likelihood ratio distortion measure, the original Itakura–Saito distortion measure is weighted in such a way that the resultant distortion has the form

$$d_{WLR} = \int_{-\pi}^{\pi} \left(S_A(\omega) - S_B(\omega)\right)\left(\log S_A(\omega) - \log S_B(\omega)\right)\frac{d\omega}{2\pi}, \tag{10}$$

where $S_A(\omega)$ and $S_B(\omega)$ are the spectra of two given models A and B. Due to the monotonicity of the logarithm function, the integrand, and hence the d_{WLR} in (10), is guaranteed to be nonnegative. When $S_A(\omega) = S_B(\omega)$ for all ω, the distortion measure is at its minimum value of zero. The weighted likelihood ratio has been applied to speech recognition in noise by Sugiyama [15] and Matsumoto and Imai [16]. They show it to be more robust to additive background wide-band noise than unweighted distortion measures—a result which is consistent with experimental results presented in this paper.

The weighted likelihood measure, like the COSH measure [4], is symmetric, i.e., $d_{WLR}(A, B) = d_{WLR}(B, A)$. Our weighted Itakura distortion, on the other hand, is asymmetric and the spectral peak weighting is applied to the test templates only. In applications where it is not feasible to train and to test in the same noisy environments, we believe that our choice of asymmetric weighting is more reasonable.

A recent paper worth mentioning describes the LPC estimation approach to AR modeling in noise by Ephraim [17]. The estimator has been used as a recognizer front-end and tested in a series of speech recognition experiments in a noisy environment [18]. Given the assumption that the noise is additive and uncorrelated with the speech signal, Ephraim formulates AR model estimation as a problem of minimizing the Itakura–Saito distortion between the received noisy spectrum and the composite spectrum of an AR speech model and the noise model. The minimization procedure is carried out iteratively. The reduction in error rate achieved by this optimal LPC estimator is equivalent to a 5 dB SNR enhancement of the received noisy speech—an improvement which is about the same as with our weighted Itakura distortion measure.

We are planning a comprehensive comparison of our method to these other methods on the same database and under identical noise conditions.

V. SUMMARY

We have proposed an adaptively weighted Itakura distortion measure and studied its effects on the performance of a conventional DTW-based speech recognizer in a series of speaker independent, isolated digit recognition experiments. Two important features of the proposed measure—1) a nonuniform spectral weighting in frequency and 2) an adaptive adjustment of the weighting factor—are experimentally confirmed to be the major factors responsible for the robust recognition performance at different SNR's. The performance of the proposed measure is the same as that of the original unweighted Itakura distortion measure at a high SNR. At medium to low SNR's, the proposed weighted measure outperforms the unweighted measure significantly. The recognition improvement is equivalent to a 5–7 dB SNR enhancement of the noise corrupted speech. Experimentally we have also shown that the estimate of segmental SNR used in our algorithm to adjust the peak weighting factor need not be very precise. No appreciable performance degradation is observed when only four brackets of segmental SNR's are used to determine the optimum weighting factor, α. The proposed weighted Itakura distortion has the same dot product form as that of the original unweighted Itakura distortion. Therefore, except for a marginal increase in computation

needed in the preparation of test templates, the complexity of the recognizer remains unchanged.

REFERENCES

[1] F. Itakura and S. Saito, "An analysis-synthesis telephony based on maximum likelihood method," in *Proc. Int. Congr. Acoust. C-5-5*, Tokyo, Japan, Aug. 1968.

[2] N. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Klatt, "Comparative study of several distortion measures for speech recognition," *Speech Commun.*, vol. 4, pp. 317-331, Dec. 1985.

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

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

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

[6] S. Kay, "Noise compensation for autoregressive spectral estimates," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 292-303, June 1980.

[7] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "The effects of selected signal processing techniques on the performance of a filter-bank-based isolated word recognizer," *Bell Syst. Tech. J.*, vol. 62, no. 5, pp. 1311-1336, May-June 1983.

[8] D. H. Klatt, "Representation of the first formant in speech recognition and in models of auditory periphery," in *Proc. Montreal Symp. Speech Recogn.*, Montreal, P.Q., Canada, June 1986, pp. 5 and 6.

[9] Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing technique in PARCOR speech analysis-synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 587-596, Dec. 1978.

[10] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. ICASSP-82*, Paris, France, May 1982, pp. 614-617.

[11] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," in *Proc. ICASSP-86*, Tokyo, Japan, Apr. 1986, pp. 85-88.

[12] J. G. Wilpon and L. R. Rabiner, "A modified *K*-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 587-594, June 1985.

[13] P. L. Chu and D. G. Messerschmidt, "A frequency weighted Itakura-Saito spectral distortion measure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 545-560, Aug. 1982.

[14] M. Sugiyama and K. Shikano, "LPC peak weighted spectral matching measures," *Electron. Commun. Japan*, vol. 64-A, no. 5, pp. 50-58, 1981.

[15] M. Sugiyama, "LPC spectral matching measures for speech recognition," Ph.D. dissertation, Tohoku Univ., Japan, Aug. 1984.

[16] H. Matsumoto and H. Imai, "Comparative study of various spectrum matching measures on noise robustness," in *Proc. ICASSP-86*, Tokyo, Japan, Apr. 1986, pp. 764-772.

[17] Y. Ephraim, "An information theoretic approach for autoregressive modeling of noisy sources," submitted for publication.

[18] Y. Ephraim, J. G. Wilpon, and L. R. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environment," in preparation.

Frank K. Soong (S'76-M'82) received the B.S., M.S., and Ph.D. degrees from the National Taiwan University, the University of Rhode Island, and Stanford University, respectively, all in electrical engineering.

Since 1982 he has been a member of the Technical Staff at AT&T Bell Labs, Murray Hill, NJ, first with the Acoustics Research Department and later with the Speech Research Department. His research work includes developing new coding algorithms at medium to low bit rates, investigating spectral distortion measures for speech processing, applying vector quantization to speaker recognition and normalization, and studying instantaneous and transitional spectral information of speech signals for speech and speaker recognition. His current interests are on segment-based speech processing techniques and their applications to speech recognition, very low bit rate speech coding, and speech synthesis.

Man Mohan Sondhi received the B.S. degree in physics (Honours) in 1950 from Delhi University, Delhi, India; the D.I.I.Sc. degree in communications engineering in 1953 from the Indian Institute of Science, Bangalore, India; the M.S. degree in electrical engineering in 1955; and the Ph.D. degree in 1957 from the University of Wisconsin, Madison.

He has been with Bell Laboratories, Murray Hill, NJ, since 1962. Before joining Bell Laboratories, he worked for $1\frac{1}{2}$ years at the Avionics Division of John Oster Mfg. Co., Racine, WI; for 1 year at the Central Electronics Research Institute of Pilani, India; and taught for 1 year at Toronto University, Toronto, Ont., Canada. At Bell Laboratories his research has included work on speech signal processing, echo cancellation, adaptive filtering, modeling of auditory, speech, and visual processing by human beings, acoustical inverse problems, and speech recognition based on hidden Markov modeling of speech. From 1971 to 1972 he was a Guest Scientist at the Royal Institute of Technology, Stockholm, Sweden.