Signal Restoration by Spectral Mapping

Biing-Hwang Juang and *L. R. Rabiner*
AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974

Abstract

Traditional approaches to the problem of noise suppression or signal restoration have been almost entirely based upon the methodology and theory of signal estimation. In this paper, we treat signal restoration as a problem in signal detection. Instead of estimating the characteristics of the signal and/or the noise, we establish a correspondence between the clean and the noisy signal through spectral mapping. In the procedure, we collect separate samples of both the clean signal and the noise. When the noise is additive, the (simulated) noisy signal is obtained by adding the noise to the clean signal. The sequence of short time spectra of the clean signal and that of the noisy signal form a one-to-one correspondence. The noisy spectral sequence is then used as a detection reference, to which the short time spectrum of an unknown noisy observation is compared, resulting in a detected occurrence of a particular group of spectra in the noisy sequence. Through the (inverse) mapping, the clean spectra that correspond to the detected noisy spectra are selected and processed to produce the restored spectrum.

One important notion of the approach is that it is not limited to the usual least squares or minimum mean square framework. Our preliminary results show that when the mapping (detection) is based upon the likelihood ratio distortion measure, an SNR improvement of approximately 10 dB is obtainable for a 14 dB SNR noisy signal. Under the same condition, an improvement of approximately 8.5 dB can be obtained using a truncated cepstral distance measure.

1.0 Introduction

Consider a signal y that is a result of contamination of a clean signal, x, by a noise signal, n. For generality, y can be written as

$$y = f(x, n) \tag{1}$$

where $f(\cdot)$ represents the contamination/distortion function that mixes the clean signal and the noise into the noisy signal y. If the contamination is additive, as it usually is, then,

$$y = x + n\ . \tag{2}$$

The general problem of interest is to restore x based upon the observation of y so that the original signal x is more faithfully represented. Traditional approaches to this problem have primarily been based upon the theory of estimation, as discussed in [1], for the case of noise contaminated speech signals.

Our attempt here, in dealing with the problem of separating x and n given y, has a more explicit focus: we try to reduce the effect of n when x, the original speech signal, is modelled as a realization of an autoregressive source. In particular, if $\sigma_x/A_x(z)$ is the optimal p-th order all-pole model of the clean speech signal x, then a processed model estimate of y, $\hat{Y}(z) = \hat{\sigma}/\hat{A}(z)$, will give, on the average, an improved similarity to the original model $\sigma_x/A_x(z)$ in terms of a predefined distortion measure d: i.e.

$$\overline{d(\sigma_x/A_x(z),\ Y(z))} > \overline{d(\sigma_x/A_x(z),\ \hat{\sigma}/\hat{A}(z))}\ , \tag{3}$$

where the overbar denotes an average and $Y(z)$ is some (spectral) estimate of the noisy signal y. The problem, formulated in a different manner (not in terms of explicit minimization of distortion measures), was discussed in [2] using maximum *a posteriori* (MAP) estimators. Our approach reported here uses (3) as the objective and takes a signal detection viewpoint in dealing with the problem. Our motivation for the distortion formulation, and thus the proposed methodology is to allow the use of sophisticated distortion measures that are believed to be more appropriate for speech signals than the usual Euclidean type, but are difficult to mathematically analyze for estimation purposes.

2.0 White Noise and Its Effects on Model Estimates

While there have been proposed quite a few distortion measures for speech signals, we focus on the likelihood ratio distortion measure and a truncated LPC cepstral distance [3] in this paper. The likelihood ratio distortion measure assumes that both of the all-pole spectra under comparison have unity gain, and is defined as

$$d_{LR}\left(\frac{1}{A}, \frac{1}{A'}\right) = \int_{-\pi}^{\pi} \frac{|A'(e^{j\omega})|^2}{|A(e^{j\omega})|^2}\,\frac{d\omega}{2\pi} - 1\ , \tag{4}$$

where $A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p}$ and $A'(z) = 1 + a_1' z^{-1} + \ldots + a_p' z^{-p}$.

The truncated cepstral distance we use is expressed as

$$d_c\left(\frac{1}{A}, \frac{1}{A'}\right) = \sum_{i=1}^{L_c} (c_i - c_i')^2 \tag{5}$$

where $\{c_i\}$ and $\{c_i'\}$ are the cepstra corresponding to the all-pole models $1/A(z)$ and $1/A'(z)$, respectively, and can be calculated from the following recursion:

$$-ic_i - ia_i = \sum_{k=1}^{i-1} (i-k)c_{i-k}a_k \quad \text{for } i > 0\ . \tag{6}$$

The length of the truncated cepstrum is denoted by L_c in (5).

In the following, let X, N and Y be the short-time spectral representations of x, n and y, respectively. The relationship among x, n and y satisfies (2). Since our focus is on the all-pole models of the signal, X, N and Y are also used as the optimal pth order all-pole spectra of x, n and y, respectively, without ambiguity. Therefore, the likelihood ratio distortion and the truncated cepstral distance between the clean and the noisy signals can be denoted by $d_{LR}(X, Y)$ and $d_c(X, Y)$. Furthermore, if $\mathbf{X}$, $\mathbf{N}$ and $\mathbf{Y}$ consist of sequences of spectral representations $\{X_t\}_{t=1}^{L}$, $\{N_t\}_{t=1}^{L}$, and $\{Y_t\}_{t=1}^{L}$ respectively, we define the average distortion/distance between any two spectral sequences by averaging the distortion/distance between corresponding spectral vectors in the sequences; for example,

$$d_{LR}(\mathbf{X}, \mathbf{Y}) = \frac{1}{L}\sum_{t=1}^{L} d_{LR}(X_t, Y_t)\ . \tag{7}$$

Let us also assume that the noise is white, Gaussian. The effects of such a noise upon the short-time signal model estimate are difficult to analyze even though they are "independent" signals. Here, we show the effects of noise upon the model estimates directly in terms of the aforementioned distortion measures. Figure 1a and 1b are plots of the distortion results $d_{LR}(\mathbf{X}, \mathbf{Y})$ and $d_c(\mathbf{X}, \mathbf{Y})$, with $p = 10$ and $L_c = 12$, as a function of the global signal-to-noise ratio (SNR). The noise is a constant power (variance) source and the SNR is defined by

$$\text{SNR} = 10\log_{10}(E_x/E_n) \tag{8}$$

where E_x and E_n are the energy of the entire signal sequence and the entire noise sequence respectively. These figures are obtained by averaging the results from 6 independent sequences of speech signals of bandwidth 4 kHz. As can be seen from the figures, when the SNR drops below 15 dB, the average distortion increases rapidly until around SNR = −15 dB. Beyond −15 dB SNR, the noise becomes dominant in the spectral representations (i.e., overtakes the formant structure in x) and the average distortion values saturate. This 30 dB range is consistent with the observation that the average spectrum of a voiced speech signal displays a −6 dB/octave trend (−12 dB/octave due to the glottal coupling and +6 dB/octave due to the lip radiation) [4], resulting in approximately a spectral dynamic range of

2368 CH2396-0/87/0000-2368 $1.00

30 dB up to 4 kHz. These figures serve as references to which results of signal restoration algorithms can be compared and calibrated.

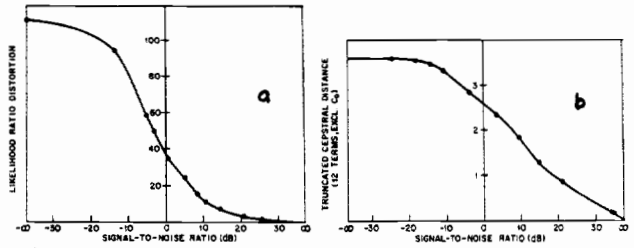

Fig. 1a) Likelihood Ratio distortion and 1b) 12-term truncated cepstral distance between the clean signal and the noisy signal as a function of the global signal-to-noise ratio.

3.0 Spectral Mapping

The spectral mapping approach we propose here tries to capitalize upon a known (a priori) correspondence between a set of clean spectra $\{X_i\}_{i=1}^{L}$ and a set of noisy spectra $\{Y_i\}_{i=1}^{L}$. The correspondence is established by adding the noise signal to the clean signal according to (2) to form the noisy signal and then calculating the spectral sets $\{X_i\}$ and $\{Y_i\}$, respectively. The most straightforward notion of spectral mapping is depicted in Fig. 2, where the clean signal space is denoted by $\mathcal{X}$ and the noisy signal space by $\mathcal{Y}$. The restoration process, for each input noisy spectrum Y, involves finding the nearest neighbor Y_i to Y in $\mathcal{Y}$ and mapping back to the clean spectrum X_i in $\mathcal{X}$, since Y_i is known to be a noisy version of X_i. Since the noise signal varies, this direct correspondence between noisy and clean spectra is only one of an infinite set of possibilities, and thus lacks the desired robustness.

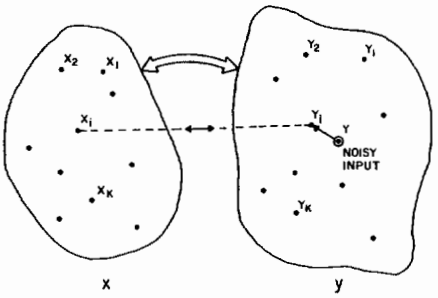

Fig. 2 Illustration of the spectral mapping scheme.

The above mapping can be easily extended to the following scheme for increased effectiveness. We assume that the $\mathcal{X}$ space consists of a finite number of signal subsources, each denoted by Z_j, $j = 1, 2, \ldots, N$. We may use the generalized Lloyd algorithm [5] to obtain all Z_j's from the clean (training) set $\{X_i\}_{i=1}^{L}$. Associated with each Z_j is a region S_j^x,

$$S_j^x = \{x \mid d(X, Z_j) \leqslant d(X, Z_i) \quad \text{for all } i\} . \tag{9}$$

Let I_j be the index set, $I_j = \{i \mid x_i \in S_j^x\}$. Since $\mathcal{X}$ maps to $\mathcal{Y}$, there exists a region S_j^y in $\mathcal{Y}$ such that $S_j^y = \{Y \mid X \in S_j^x\}$ and obviously $Y_i \in S_j^y$, for $i \in I_j$. We now define a modified distortion measure in $\mathcal{Y}$, from a noisy spectrum Y to a noisy region S_j^y as depicted in Fig. 3,

$$d'(Y, S_j^y) = \frac{1}{\|I_j\|} \sum_{i \in I_j} d(Y, Y_i) \tag{10}$$

where $\|\cdot\|$ denotes the cardinality. The nearest neighbor of an input Y is the region S_j^y that satisfies

$$d'(Y, S_j^y) \leqslant d'(Y, S_i^y) \quad \text{for all } i . \tag{11}$$

When (11) is true, we say Z_j is the restoration spectrum of the 1-nearest-neighbor choice.

Conversely, we define the region $U(Y) \subset \mathcal{Y}$ to be the set $\{Y' \mid Y' \in \mathcal{Y},\ d(Y, Y') \leqslant d_t\}$. Represented in terms of the given noisy set $\{Y_i\}_{i=1}^{L}$, $U(Y) \supset \{Y_i \mid d(Y, Y_i) \leqslant d_t\} \triangleq U^*(Y)$. Note that

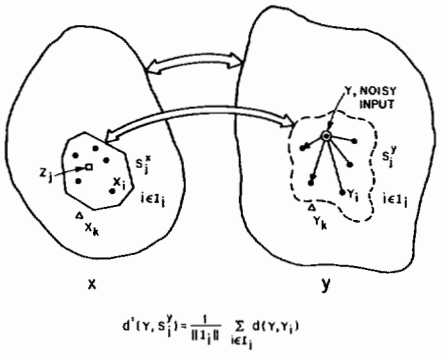

Fig. 3 Illustration of the spectral mapping scheme with a modified distortion/distance measure.

$U(Y) = \mathcal{Y}$ if $d_t = \infty$ for any Y. The notion of $U^*(Y)$ is of practical importance in the current approach which relies upon the given training set for the representation of the entire space. The distortion threshold, d_t, allows adjustment of the subspace in $\mathcal{Y}$, within which other processing can be performed. When d_t is finite, the modified distortion of (10) can be further revised to reflect the subspace processing; in particular, we define

$$d''(Y, S_j^y; U^*(Y)) = \frac{1}{\|I_j^*\|} \sum_{i \in I_j^*} d(Y, Y_i) \tag{12}$$

where $I_j^* = \{i \mid X_i \in S_j^x \text{ and } Y_i \in U^*(Y)\}$. If $I_j^* = \{\text{empty set}\}$, $d''(Y, S_j^y; U^*(Y)) \triangleq \infty$. By ordering the modified distortion d'', one can then retrieve the appropriate clean subsource representations Z_j for restoration. Similar to (11), for example, we may choose Z_j when $d''(Y, S_j^y; U^*(Y)) \leqslant d''(Y, S_i^y; U^*(Y))$ for all i as the 1-nearest-neighbor choice. Alternatively, we may define

$$I''(Y; d_t'') = \{j \mid d''(Y, S_j^y; U^*(Y)) \leqslant d_t''\} \tag{13}$$

and obtain the restoration vector by

$$\hat{X}(Y) = \frac{1}{\|I''(Y; d_t'')\|} \sum_{i \in I''(Y; d_t'')} Z_i . \tag{14}$$

3.1 Practical Implementation

The system we implemented is based upon the modified distortion measure of (12). The distortion threshold d_t, in the definition of $U^*(Y)$, can be a fixed value or a fluctuating number. In the latter case, d_t can be chosen so that $\|U^*(Y)\|$ is a constant N_b. In our preliminary experiments, the difference in the spectral distortion which resulted from the restoration process using either a fixed threshold or a constant $\|U^*(Y)\|$ was not apparent. We therefore chose to use the case with constant $\|U^*(Y)\|$. Furthermore, we use (13) and (14) to obtain the restored spectrum. Similar to the situation in defining $U^*(Y)$, we chose to use fluctuating d''_t in (13) such that $\|I''(Y; d''_t\|| = N_a$ where N_a is a constant for ease in implementation. We varied N_a and N_b in our experiments to study their effects, the results of which are reported in Section 4.

The distortion measures used, as mentioned above, are the likelihood ratio distortion and the truncated cepstral distance. When the likelihood ratio measure is used, the averaging procedure in (14) is performed on the residual-normalized autocorrelation vectors. When the cepstral measure is used, the averaging is on the cepstral vectors directly. This is clear from a distortion-minimization point of view.

4.0 Experimental Results

4.1 Database

The speech material we collected for the training sequence was recorded from 20 speakers, 15 male and 5 female, each speaking 5 sentences, resulting in a total of 100 different sentences with an accumulated duration of about 6 minutes. The test material, similarly, consisted of 6 sentences spoken by two speakers, who were not used to

provide the training sequence. These sentences were extemporary and conversational; in addition, each sentence had different content. The reason for choosing a database of such generality was that our study tried to concentrate on the feasibility of the new approach under the worst speaking conditions.

All the speech material were analyzed using a 10th order LPC autocorrelation method. The analysis window was 20 msec long and the frame rate was 80 per second (i.e. 12.5 msec shift). As a result, the total numbers of vectors used as the training sequence and the testing sequence, respectively, were 27310 and 1562. We did not try to process or reduce the training database for higher efficiency.

The noise data were generated using a standard pseudorandom number generator with Gaussian distribution. The noise used in simulating the noisy signals was different for all the sentences so as to maximize the true statistical noise perturbation upon the signal. The noise contamination in our simulation is of a global type in that the noise power is kept constant. This, of course, does not allow evaluation of the approach under a constant signal-to-noise ratio condition. The global SNR therefore represents an average over all the sentences. In our current study, the global SNR was approximately 14 dB overall.

4.2 The Likelihood Ratio Measure

After the establishment of the clean and noisy sequences, and the correspondence between the two, the next step in the approach is to generate N representative vectors $\{Z_j\}_{j=1}^{N}$, each being associated with a region S_j^x, in $\mathcal{X}$. We used the generalized Lloyd method for this step. In our experiment, we tested two values of N, namely $N = 256$ and $N = 64$. These two conditions will be referred to as 8-bit VQ and 6-bit VQ. As stated in Section 3.1, $\|U^*(Y)\|$ was kept constant in the experiment. However, we tested different $\|U^*(Y)\|$ values in combination with the 6-bit and 8-bit VQ cases to study its effect.

Figure 4 shows the average likelihood ratio distortion between the original clean spectra and the restored spectra as a function of the number of nearest neighbors for averaging, expressed in logarithmic units. The untreated noisy spectral sequence, which has $\sim$14 dB SNR, results in an average likelihood ratio distortion of 6.8 as shown by the dashed line in Fig. 4. The six curves are designated by the six parenthesized pairs respectively. The first number in each pair is $\|U^*(Y)\|$, the noisy locality number, and the second designates 6 or 8-bit VQ codebook. The order of these pairs in the figure is according to the resultant distortion of each case at $N_a = 1$ (or $\log_2 N_a = 0$). Several observations can be made from the results of Fig. 4. First, the treated/restored all-pole spectral sequences produce much lower distortions than the noisy, untreated all-pole sequence. Second, the distortion decreases almost monotonically with increases in N_a; however, the reduction in distortion is insignificant when N_a increases past 8 or 12. Third, better results are consistently obtained with smaller $\|U^*(Y)\|$; at $N_a = 1$ and 2, the distortion results are ordered exactly according to the $\|U^*(Y)\|$ number — larger $\|U^*(Y)\|$ produces higher distortion — for each VQ case. This result confirms and validates the original motivation of the approach in that the spectral perturbation due to noise can be somehow traced (via the known mapping) and processed simply by the nearest neighbor principle. When N_a is more than 4, cases with large $\|U^*(Y)\|$ continue to improve while cases with small $\|U^*(Y)\|$ start showing plateauing in performance. Finally, with the same noisy locality number $\|U^*(Y)\|$, 8-bit VQ produced lower distortion than 6-bit VQ.

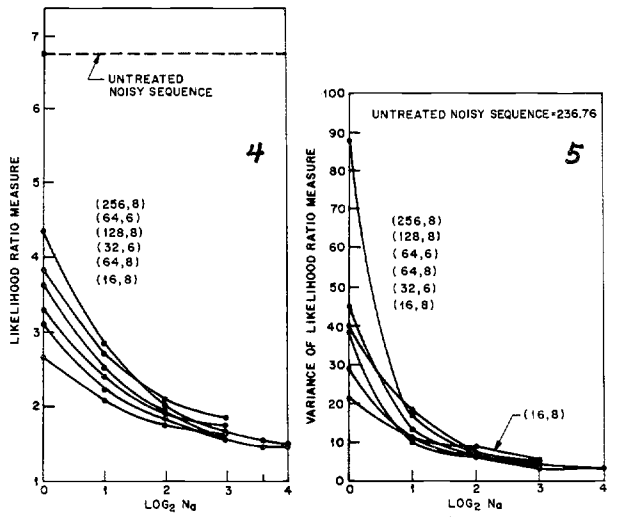

Fig. 4 Average likelihood ratio distortion and 5) average variance of the distortion of the restored sequences as a function of N_a, for several sets of system parameters.

The above best result, when plotted against Fig. 1a of the untreated noisy sequence, shows an effective improvement of 10 dB SNR.

Another important performance indicator is the average variance of the resultant distortion, which is plotted in Fig. 5. Higher distortion variance means more uneven distortion, a situation generally undesirable in most speech processing techniques. The distortion variances produced by the treated spectral sequences are significantly lower than the variance of the untreated case (236.76). This reduction in average variance alone is as, or even more, important than the reduction in the average distortion. The relationship among the distortion variance, the noisy locality number $\|U^*(Y)\|$ and the averaging number N_a, however, is not as clear as that for the average distortion as shown in Fig. 4.

4.3 Truncated Cepstral Distortion Measure

We repeated the above experiment using a 12-term truncated cepstral distance for noisy spectral mapping. This 12-term truncated cepstral distance does not include the zeroth cepstral coefficient which corresponds to the logarithm of the gain term of the all-pole spectrum. The average cepstral distance is plotted in Fig. 6, again as a function of $\log_2 N_a$, for various $\|U^*(Y)\|$ and VQ combinations. The general observations are quite similar to those already given for the likelihood ratio distortion case. Note that the cepstral distance is a Euclidean distance and thus symmetric. This has one advantage in that the measure and the mapping direction can be arbitrary. Nevertheless, the measure is not consistent with the all-pole model measure; that is, being close in the truncated cepstral distance sense need not mean close in the LPC model spectrum sense. This accounts for the increase in the average truncated cepstral distance for the case of (64, 6) at $N_a = 1$. The point to note here, however, is that the nearest neighbor

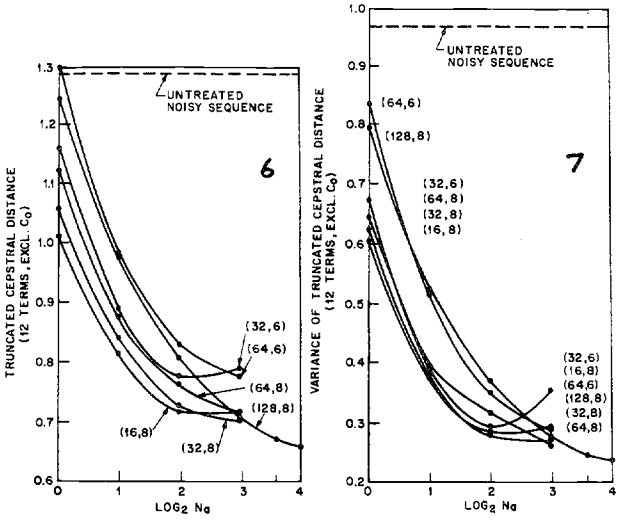

Fig. 6 Average truncated (12-term excluding c_0) cepstral distance and 7) average variance of the distance of the restored sequences as a function of N_a, for several sets of system parameters.

mapping is still valid for a Euclidean distance even though the measure may not be consistent with the LPC model analysis criterion. Indeed, Fig. 6 shows that a significant reduction in the average truncated cepstral distance is obtainable with the current restoration procedure based upon spectral mapping.

While the overall results of Fig. 6 resemble those of the likelihood ratio measure case (Fig. 4), we observe one extra phenomenon in the truncated cepstral distance case. For the case of (32, 6) and (16, 8), i.e. cases with coarse VQ or tight noisy locality, the average cepstral distance increases when N_a increases above 4. This is caused by excessive averaging within the available noisy locality. This increase in average distance, however, is small compared to the reduction obtained from the average distance of the untreated sequence. In fact, the above best result, when plotted against Fig. 1b of the untreated noisy sequence, shows an effective improvement of $\sim$8.5 dB SNR.

The resultant variances of the truncated cepstral distance are plotted in Fig. 7 for various cases. The minimum average variance obtained was about 0.24, which is significantly smaller than the average variance of the untreated sequence (0.97).

4.4 Truncated Cepstral Measure with Gain Term

Although our main interest was the restoration of the all-pole spectral shape, we also investigated the recovery of the signal level in terms of the LPC model gain using the truncated cepstral distance. We used a 13-term cepstral distance, adding the squared difference of the logarithm of the gain term to the 12 term truncated cepstral distance described above. The reason for using this truncated cepstral measure for the spectral level recovery study is simply because of the Euclidean properties of the distance measure.

The same experiment reported above was repeated using the 13-term truncated cepstral measure. The resultant average cepstral distance is plotted in Fig. 8. Compared to the untreated noisy sequence which produced an average distance of 11.76, the results of various parameter combinations show remarkable improvements. We also observe significant reduction in the average cepstral distance variance as shown in Fig. 9. The untreated noisy sequence produced an average distance variance of 186.5, about 15-20 times that of the restored results with the best parameter settings.

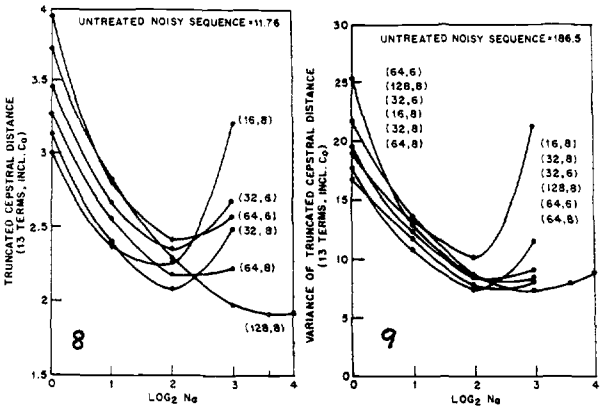

Fig. 8 Average truncated (13-term including c_0) cepstral distance and 9) average variance of the distance of the restored sequences as a function of N_a, for several sets of system parameters.

4.5 Other Results

The above results are based upon the comparison of the average spectral distortion. The effectiveness of the proposed approach can be further illustrated by comparing the resultant distortion sequences which not only show the average distortion improvement but also show the fine sequential details of the distortion for each frame of the signal. The distortion sequences are plotted in Fig. 10 for the likelihood ratio measure case. The first row of each figure is the energy contour of the test signal. The sequence has 161 frames. The second row is the distortion sequence obtained by comparing the clean spectral sequence and the untreated noisy spectral sequence. As can be seen, there is a strong inverse correlation between the energy and the untreated distortion sequence: signal sections with higher energy, in general, are less affected by the noise and thus have lower distortion than lower energy sections. However, depending on the signal characteristics, some sections with sufficiently high energy may still be seriously distorted by the noise contamination. This is especially evident at phonological transitions of the signal. The 3rd, 4th and 5th rows are the distortion sequences for $N_a = 4$, 8 and 16 respectively. The reduction in the average spectral distortion is obvious in these plots. Furthermore, it can be seen that the mapping/averaging procedure eliminated all of the extremely high distortion regions of the untreated sequence. Many of these regions correspond to the transitional or low energy sections of the signal. The effects of averaging are also demonstrated in these figures.

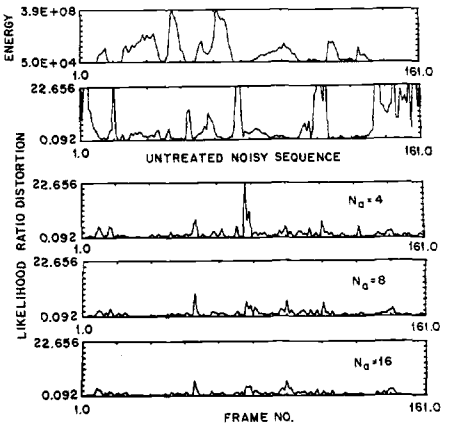

Fig. 10 Distortion sequence of the noisy and the restored spectra when using a likelihood ratio distortion measure.

References

[1] J. S. Lim, *Speech Enhancement*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[2] J. S. Lim and A. V. Oppenheim, "All-pole Modelling of Degraded Speech," in *Speech Enhancement* ed. J. S. Lim, p. 101-114, Prentice-Hall, New Jersey, 1983.

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

[4] H. Wakita, *Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic/Articulartory Conversion Methods*, SCRL Monograph No. 9, Speech Communications Research Laboratory, Santa Barbara, CA 1972.

[5] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization," IEEE Trans. Comm. vol. COM-28, pp. 84-95, Jan. 1980.