CMU-CS-78-124

COMPARISON OF SPEECH SPECTRA FOR ADDITIVE TYPE OF SPECTRAL DISTORTION

B. Yegnanarayana and D. Raj Reddy
Department of Computer Science
Carnegie-Mellon Univesity
Pittsburgh, PA 15213

May, 1978

This work was supported by the Defense Advanced Research Projects Agency (F44620-73-C-0074) and is monitored by the Air Force Office of Scientific Research.

ABSTRACT

Current methods of measuring dissimilarity between spectra using linear predictor coefficients or cepstral coefficients are shown to be deficient in many ways. In particular we show that certain types of degradation in speech data can significantly affect these parameters making them essentially useless for comparison of spectra. Two types of degradation namely, the quantization noise of waveform encoding (ADPCM) and additive band-limited Gaussian noise, are considered for illustration. For these two cases the linear predictor coefficients do not represent the true envelope of the short time spectrum. Earlier studies reported large values of distance between ADPCM and original data. Those values are largely due to differences in the spectral dynamic range. By using a new measure proposed in this paper, we show that the true differences in spectral envelopes of ADPCM and original data are insignificant. This result also explains to some extent the good recognition capability of Harpy continuous speech recognition system for ADPCM data even for the lowest bit rate. Based on this new measure we arrive at the conclusion that the effect of additive noise on parametric extraction is more severe than quantization noise.

1. INTRODUCTION

Parameters representing smoothed spectral characteristics of short segments of speech are often used as features in several speech processing systems [1]. Envelope of short time spectrum represents the shape of the vocal tract during the analysis period and temporal variation of the envelope is supposed to contain speech and speaker information. Several parameter sets are available to describe the envelope characteristics of short time spectrum. Autocorrelation coefficients, which are the Fourier coefficients of short time spectrum, are the basic parameter set from which most of the other parameters are derived. Inverse Fourier transform of truncated autocorrelation coefficient series gives the autocorrelation smoothed spectrum. A more accurate estimation of spectral envelope, especially at its peaks, is through linear prediction coefficients (LPC) which are derived from autocorrelation coefficients by solving a set of normal equations [2]. The physical basis of LP (Linear Prediction) smoothing is that it gives the frequency response of the best all pole model for vocal tract system. Other useful parameters are cepstral coefficients, reflection coefficients, area coefficients and formants, all of which can be related to LPC or autocorrelation coefficients [2].

The main problem in pattern recognition in speech involves matching the test spectrum with a reference spectrum using a suitable

distance measure. The reasons for the choice of smoothed spectrum for comparison are: (1) Fewer parameters are needed to describe the spectral behavior compared to the actual waveform or its spectral values, resulting in large data reduction and saving in computation time. (2) The smoothed spectrum contains most of the information needed to desrcibe the speech signal as evidenced by the perceptual studies in analysis-synthesis telephony.

Several possibilities exist for matching the spectra [3]. Distance measures directly on the parameter sets have not yielded good results in speaker and speech recognition studies [4], [5]. But high scores were reported [4] in a speaker identification test while using a root mean square (rms) Euclidean distance measure between test and reference log spectra from LP analysis. This is refered to as an rms log spectral measure. Atal [6] has shown for a speaker verification test that weighted Euclidean distance measure based on cepstral coefficients resulted in the highest scores among several parameter sets. The unweighted Euclidean distance based on cepstral coefficients is referred to as a cepstral distance measure. Magill [7] and Itakura [8] have proposed the ratio of LP residual energies for comparing reference and test data. Itakura [8] has shown that logarithm of the ratio of LP residuals (log likelihood ratio) resulted in high recognition scores in a word recognition experiment. Gray and Markel [3] proposed a cosh measure which is obtained by averaging two nonsymmetrical likelihood ratios. This new measure

possesses the desirable symmetry property of a distance measure although the individual log likelihood ratios do not. Interrelationships between the different measures were studied and it was found that cepstral measure and cosh measure satisfy the general criteria useful for measures of distance in speech procesing [3].

The mathematical distance measures proposed so far have been found to show no correspondence with the perceptual data [3]. Perceptual experiments did not show a unique value for a distance measure that corresponds to a barely perceptible change in formant frequency over a range of formants and frequencies. But a relation between distance measure and perceptual changes would help in fixing threshold value for the distance to determine significant changes in data. The conclusion of Gray and Markel sutdy on distance measures is that "until perceptual experiments or speech recognition tasks, for example, show some other distance measure to be more meaningful for speech processing, the rms log spectral distance makes the best reference point for comparision. It can be physically interpreted; it is analytically tractable, easily and efficiently computed (using the cepstral measure), and relatable to several other widely used measures of distance."

It is interesting to note that the starting point for all the distance measures proposed so far is the LP smoothed spectrum. It is implicitly assumed that any changes in the original spectrum are reflected faithfully in the LP spectrum as well. Several changes in

the original speech spectrum may occur as a result of different sources of variability in speech input, but these changes may not produce perceptually different sounds. However, the variability in the original spectrum may produce large variation in the perceptually significant features of smoothed spectrum as a result of the transformation involved in generating the smoothed spectrum. The very fact that several distance measures are interrelated and all aim at comparing the smoothed log spectra, shows that they have no direct relation to the differences present in the actual spectra.

In this paper we shall investigate how the actual differences in the original spectra are reflected in various distance measures, especially in the most commonly used LP distance proposed by Itakura [8]. We shall discuss the inadequacy of the measure particularly while comparing an original speech data with its distorted version. We shall consider two types of distortions for illustration: quantization noise of waveform encoding schemes and additive band-limited Gaussian noise. A more practical approach for comparing such data is proposed.

II. DISCUSSION OF THE PROBLEM

In this section we shall discuss the following problem:

Suppose we have two spectra whose envelopes are to be compared. What is a suitable measure for comparison? How good are the smoothed spectra obtained by linear prediction analysis or cepstrum analysis for such a comparison?

Before we discuss these questions let us consider a few exmples to explain the problem. Consider the case of comparison of two spectra having the same spectral envelope but differing in their absolute values (not on log scale) by a constant (K). That is if $P_1(\omega)$ = $P(\omega)$ then $P_2(\omega)$ = $P(\ )+K$. Or, the two spectra may have similar envelopes but they differ in their average slopes (again not on log scale). A slightly more complex case could be one in which the spectral envelopes may differ in slopes only in certain frequency ranges. Such changes in envelopes may occur due to several distortions the speech signal undergoes before actual processing for parameter or feature extraction. As for example, small amounts of quantization noise of waveform encoded speech or additive background noise generally produce insignificant changes in the spectral envelope. As another example, the response characteristics of the antialiasing filter are superimposed on the speech spectrum thus contributing to changes in its envelope. There are several other

sources such as the frequency response of input transducer, telephone distortions etc. which alter the shape of the spectrum.

It is evident that basically there are two types of spectral distortions: additive and multiplicative. In the case of additive uncorrelated noise the spectral envelope of noisy speech will be sum of spectral envelopes of speech and noise. If the signal to noise ratio (SNR) is high, then the spectral envelope of speech is practically unaffected by the noise. In the case of superimposed frequency responses of filter or transducer characteristics, the spectral envelope of the distorted speech is a multiplication of the frequency response and the spectal envelope of the undistorted speech. When we consider log spectrum, the envelope is altered with the addition of noisy spectrum to the original even though the SNR is high. The reason for this is that the spectral dynamic range is significantly altered with the additive term in the spectrum. On the other hand, in the case of multiplicative spectral distortion, the smoothed log spectrum is just an addition of the smoothed log spectra of the frequency response of the transducer and the envelope of the original spectrum. In the latter case it is possible to subtract the superimposed log frequency response to obtain the original smoothed spectrum. Such a simple scheme does not work for additive noise distortion.

The implication of the above discussion in the two familiar methods of spectral smoothing viz., cepstrum and linear prediction, will be considered now.

a) Cepstrum Analysis

Let

$P(\omega)$ = Spectrum of speech segment

$N(\omega)$ = Spectrum of noise

$\{c(n)\}$ = Cepstral coefficients

$\{R(n)\}$ = Autocorrelation coefficients

Then, using Fourier series expansion of power spectrum and the definition of cepstrum we get

$$P(\ ) + N(\ ) = R(0) + \sum_{k=1}^{\infty} R(k) \cos(k\omega) \qquad (1)$$

and

$$c(0) + \sum_{k=1}^{\infty} c(k) \cos(k\omega) = \ln [\, P(\omega) + N(\omega) \,]$$
$$= \ln [R(0) + R(k) \cos(k\omega)] \qquad (2)$$

Cepstrally smoothed spectrum [9] is obtained by finding the frequency response of the truncated series in the summation on the LHS of (2). It is obvious that the cepstral coefficients and hence the resulting smoothed spectrum are strongly dependent on $N(\omega)$ through $R(n)$. The important point is that the dynamic range of $P(\omega)$ is reduced affecting the envelope of log spectrum and hence cepstrum. The effect of noise is also sometimes reflected as spurious peaks or wide bandwidth formants in the smoothed spectrum. Fig. 1 shows cepstrally smoothed spectra for a voiced speech segment and for additive and multiplicative types of spectral distortions in the segment. The frequency response used for generating the multiplicative distorted speech is given in Fig. 2.

b) LP Analysis

Similar effects can be observed in LP smoothed spectrum also as shown by the following analysis. The approximate spectrum $\hat{P}(\omega)$ in LP analysis is derived by minimizing the integrated ratio of the original $P(\omega)$ and the approximate spectrum [2]. That is $\hat{P}(\omega)$ minimizes the function

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)}\, d\omega \qquad (3)$$

subjected to the condition

$$\int_{-\pi}^{\pi} P(\omega)\, d\omega = \int_{-\pi}^{\pi} \hat{P}(\omega)\, d\omega \qquad (4)$$

for undistorted speech, and it minimizes the function

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(\omega) + N(\omega)}{P(\omega)}\, d\omega \qquad (5)$$

subjected to the condition

$$\int_{-\pi}^{\pi} [P(\omega) + N(\omega)]\, d\omega = \int_{-\pi}^{\pi} P(\omega)\, d\omega \qquad (6)$$

for speech corrupted with additive noise. Fig. 3 illustrates the effect of the two types of distortion considered in Fig. 1 on LP smoothed spectrum. It is clear that additive noise reduces the dynamic range and widens the bandwidth of some formants. Formant estimation from multiplicative type spectral distortion is not

significantly affected unless the superimposed frequency response filters out certain frequency components.

The best way to illustrate the effect of addiive noise on LP smoothing is by considering smoothed spectra obtained by computing LPCs for autocorrelation coefficient sets which differ only in the value of the zeroth coefficient R(0). Fig. 4 shows that as R(0) is increased the dynamic range of the spectrum is reduced and the formant bandwidths are increased. The higher formants do not appear in the LP spectrum when R(0) is increased, although they are present in the actual spectrum since only R(0) is altered.

While comparing spectra in a speech processing system the reference features are collected on a reasonably clean data whereas the test data may be subjected to various types of distortions. It is not possible to collect and store references for distorted speech, especially for additive noise type of distortion. This is because the source of variability is not known in advance. Normalization for multiplicative type of distortion like transducer or filter frequency response can be made. But for additive type of distortion one may end up in comparing LP spectra (a) and (d) in Fig. 4 and conclude that they are significantly different although the envelope information is unaffected in their actual spectra.

The above discussion suggests that a different approach for comparing speech spectrum envelopes is needed in order to overcome the difficulties indicated. The problem arises mainly because the

comparison is made on envelopes derived from the original spectra through transformations which do not preserve all the features of the original envelope. We shall develop a method in the following section which brings the two spectra to the same level of dynamic range before comparing them. We shall use this method to show that in most cases the large distance measured between original and distorted speech spectra for additive noise distortion is mainly due to differences in their dynamic ranges and not due to differences in their spectral shapes [10]. The result is illustrated for quantization and additive white noise distortions.

III. PROPOSED METHOD

Since our objective is to compare shapes of spectral envelopes, it is important to determine the parameters that represent the envelope information. Autocorrelation coefficients are the coefficients of Fourier series expansion of power spectrum, in which the zeroth coefficient merely determines the average value of the spectrum. All the other coefficients determine the shape of the spectrum. In LP smoothing the influence of the zeroth autocorrelation coefficient is significant, as shown in Fig. 4, making comparison of smoothed spectra difficult.

To overcome this difficulty we propose the following scheme for comparison of two spectra. It consists of modifying one autocorrelation set relative to the other before using a smoothing

technique. Let $\{R_1(i)\}$ and $\{R_2(i)\}$, $i=0,1,2,..M$ be the autocorrelation sets of the original and distorted speech respectively. In order to keep the effect of the zeroth coefficient same in both cases, we multiply the coefficients $R_1(i)$, $i=1,2,...M$ by a constant K which is determined by minimizing the error function

$$E = \sum_{i=1}^{M} [K r_1(i) - r_2(i)]^2 . \qquad (7)$$

$dE/dK = 0$ yields

$$K = \frac{\sum_{i=1}^{M} r_1(i)\, r_2(i)}{\sum_{i=1}^{M} r_1^2(i)} . \qquad (8)$$

where $r_1(i)=R_1(i)/R_1(0)$ and $r_2(i)=R_2(i)/R_2(0)$.

The modified autocorrelation set consists of $R_1(0)$, $KR_1(1)$, $KR_1(2),...KR_1(M)$ and let us denote this set as $\{R(i)\}$, $i=0,1,...M$. Notice that the spectral shape will not be altered due to this

modification. The LP smoothed spectra for $R_1(i)$ and $R(i)$ will however be different. It appears logical to compare the smoothed spectra derived from $R(i)$ and $R_2(i)$ using any of the standard distance measures [3]. If the spectral envelopes of the two spectra are same, any distance measure operated on the LP spectra should yield zero value.

For illustration we shall use the log likelihood ratio measure proposed by Itakura [8] for comparing the smoothed spectra. This measure is computationally more efficient than the rms log spectral measure and it has been effectively used in problems of speech recognition [8], speaker recognition [11], and variable frame rate synthesis [12], although from statistical considerations the measure was found to be unsatisfactory [13]. Let $\{r(i)\}$, $\{r_1(i)\}$ and $\{r_2(i)\}$ be the normalized autocorrelation coefficients and $\{a(i)\}$ $\{a1(i)\}$ and $\{a2(i)\}$ be the linear prediction coefficients corresponding to $\{R(i)\}$, $\{R_1(i)\}$ and $\{R_2(i)\}$ respectively. The measure developed by Itakura for autocorrelation method of linear prediction to evaluate the dissimilarity between original and test segments is given by [10]

$$d = \ln \left(AVA^T / BVB^T \right) \qquad (9)$$

where the vectors A and B are the augmented LPC vectors of the original and test segments respectively, i.e.,

$A = (1, a(1), a(2), \ldots a(M))$

and

$B = (1, b(1), b(2), \ldots b(M))$.

V is the autocorrelation matrix of the test segment. To take into account the length of the speech segment it is preferable to consider as distance measure

$$D = N_{eff} \cdot d \qquad (10)$$

where N_{eff} denotes the effective length of the segment in samples. If the segment is multiplied by a Hamming window, $N_{eff} = 0.3975N$ where N is the actual number of speech samples in the segment. For a rectangular window $N_{eff} = N$.

An efficient method of obtaining d is by computing the log likelihood ratio from the residual energies δ and α as follows [3].

$$d = \ln (\delta/\alpha) \qquad (11)$$

where

$$\delta = AVA^{T} = \sum_{l=-M}^{M} r_a(n)\, r_x(n), \qquad (12)$$

$$\alpha = BVB^{T} = \sum_{i=-M}^{M} r_b(n)\, r_x(n) \quad . \qquad (13)$$

$\{r_x(n)\}$ is the normalized autocorrelation sequence of the test data

and $\{r_a(n)\}$ and $r_b(n)\}$ are the autocorrelation sequences of $\{a(i)\}$

and $\{b(i)\}$ respectively.

In the illustrations to follow we shall compute two distances D_1 and D_2 for each frame of speech data, given by

$$D_1 = N_{eff} \cdot \ln (\delta_1/\alpha_1) \qquad (14)$$

and

$$D_2 = N_{eff} \cdot \ln (\delta_2/\alpha_2) \qquad (15)$$

where

$$\delta_1 = \sum_{i=-M}^{M} r_{a1}(n)\, r_2(n) \quad , \qquad (16)$$

$$\alpha_1 = \sum_{i=-M}^{M} r_{a2}(n)\, r_2(n) \quad , \qquad (17)$$

$$\delta_2 = \sum_{i=-M}^{M} r_a(n)\, r_2(n) \quad , \qquad (18)$$

and

$$\alpha_2 = \alpha_1 \quad . \tag{19}$$

$\{r_a(n)\}$, $\{r_{a1}(n)\}$ and $\{r_{a2}(n)\}$ are autocorrelations of the LPCs $\{a(i)\}$, $\{a1(i)\}$ and $\{a2(i)\}$ respectively, defined earlier. The distance D_1 is a measure of dissimilarity between original and distorted speech segments and the distance D_2 is a measure of dissimilarity between the same segments after the modification of the autocorrelation coefficients suggested in equations (7) and (8).

In most cases of additive type of distortion the normalized autocorrelation coefficients of the distorted segments are usually smaller than the corresponding coefficients of the original data. Therefore in most cases the parameter K defined in Equation (8) is less than unity. But for some segments of speech the value of K may be greater than unity while comparing with their distorted versions. This happens generaly for segments having significant high frequency energy. In such cases K is forced to be unity, i.e., no modification will be done for the autocorrelation coefficients of the original data. The value of K may also exceed unity while comparing two segments corresponding to different sounds as in a speech recognition experiment. In such a case the roles of $R_1(i)$ and $R_2(i)$ are to be reversed in (8) so that we get a value of K less than unity. Also,

the values of $R_2(i)$, i=1,2,...M are to be multiplied with the new K.

In other words $R_2(i)$ will be modified with respect to $R_1(i)$.

It may be tempting to carry out the modification always on one set of autocorrelation irrespective of the value of K, but for K>1 the modification results in increasing spectral spread or dynamic range of the smoothed envelope. This is not desirable for two reasons. Firstly, increasing dynamic range results in an unstable all-pole filter. Secondly, even small differences in formant frequencies and bandwidths yield large values of distance

While comparing spectra for different sounds, it is preferable to adopt two-way modification of the autocorrelation sets. This is done as follows. First the value of K is computed as

$$K = \frac{\sum_{i=1}^{M} r_1(i)\, r_2(i)}{\sum_{i=1}^{M} r_1^2(i)} \quad . \qquad (20)$$

If this K is less than or equal to one, then $\{R_1(i)\}$ is modified as

$R_1(0)$, $KR_1(i)$, $KR_1(2)$,... $KR_1(M)$. If this K is greater than one, then a new value of K is computed as

$$K = \frac{\sum_{i=1}^{M} r_1(i)\, r_2(i)}{\sum_{i=1}^{M} r_2^2(i)} \quad . \qquad (21)$$

The new K will be less than one. Therefore, the set $\{R_2(i)\}$ is modified as $R_2(0), KR_2(1), KR_2(2), \ldots KR_2(M)$ and compared with the set $\{R_1(i)\}$. This two way modification assures the stability of the modified all-pole filter and also the two spectra being compared are brought to the same lower level of dynamic range irrespective of the order in which they are considered for computing K. It is to be noted that for $K<0$ no modification should be done because the spectral shape will be changed.

IV. GENERATION OF DISTORTED SPEECH

A. Quantization Noise

As an illustration of the application of the proposed method for comparing spectra, we consider two types of distortions in speech: quantization noise of an adaptive differential pulse code modulation (ADPCM) and additive white Gaussian noise. In order to compare our results with those obtained by Sambur and Jayant [10], we use their method for generating ADPCM speech. The scheme, shown in Fig. 5, uses a forward adaptive quantization and time invariant first

order predictor. The optimum quantization step Δ_{opt} is computed from the variance for a block of N input samples and it is updated for quantization of every new block. The following equations define the differential coding:

X_n = Input error samples

E_n = Prediction error samples

X_{nq} = Quantized input speech samples

E_{nq} = Quantized error samples

F = Sampling frequency (kHz)

B = Number of bits per sample

I = Bit rate (kilobits per sec) = F.B

SQNR = Signal to quantization error ratio

$$= \sum_n X_n^2 \Big/ \sum_n (X_n - X_{nq})^2$$

$$E_n = X_n - A_1 X_{(n-1)q}$$

$$X_{nq} = A_1 X_{(n-1)q} + E_{nq}$$

$$\Delta_{opt} = K_{opt} \left[\sum_{n=2}^{N} (X_n - A_1 . X_{n-1})^2 / (N-1) \right]^{1/2}$$

where A_1 = 0.875 and K_{opt} for different values of B are given in Table 1. The value of N was chosen to be 64.

TABLE 1. Design values of K_{opt} for different values of B [10]

B	2	3	4	5	6
K_{opt}	0.996	0.586	0.335	0.225	0.120

B. Additive White Noise

Sequences of independent Gaussian noise samples, band-limited to about 4 kHz, are added to speech samples X_n to produce data with additive noise distortion. The signal to additive noise ratio (SANR) is given by

$$SANR = \sum_{n=1}^{N_T} [X_n^2 / (\sigma^2 . N_T)] \qquad (22)$$

where N_T is the total number of samples in the input for a given utterance. The variance σ^2 of the noise sequence is varied to produce speech data with different signal to noise ratios. The SANR values for the test utterance chosen in this study are 10, 14.6, 19.2, 22.5 and 27.0 dB corresponding to SQNR values of ADPCM data for B = 2,3,4,5 and 6 respectively.

V. COMPARISON OF ORIGINAL AND DISTORTED SPEECH DATA

The sentence "DO ANY PAPERS CITE NILSSON" spoken by a male speaker into a close speaking microphone was used as speech data in this study. The signal was prefiltered (85-4500 Hz) and sampled at 10 kHz. The samples were stored as 9 bit numbers. Distorted speech was generated using the methods described in Sec.IV. Frame by frame analysis was performed to compare the original and the distorted speech data. The number of samples per frame was chosen to be 200 corresponding to 20 msec of speech segment. The data was multiplied with a Hamming window and was passed through a pre-emphasis filter $(1-0.92z^{-1})$ before computing the autocorrelation coefficients. The coefficients of a 14 pole all-pole filter were obtained by solving the autocorrelation normal equations [2]. The distances D_1 and D_2 defined in (14) and (15) as measures of dissimilarity between original and distorted speech data were computed for all the 79 frames in the test sentence. The value of N_{eff} in D_1 and D_2 is 80.

It is important to note the main difference in the nature of the two types of distortions. Although the signal to noise ratio (SNR) for the complete utterance is same in both cases, the SNR for individual frames are different as shown in Figs. 6 and 7. The shapes of SNR contours are same for different bit rates of ADPCM data and also for different total SNR of additive noise degradation.

However, the variation in SNR for quantization noise is much smaller than the variation for additive noise. The variation itself is due to different values of signal energy in different frames. This information together with the spectral distribution of noise in each frame should be considered while interpreting the measure of dissimilarity due to distortion.

The LP distance contours measuring the dissimilarity between original and distorted speech are shown in Figs. 8 and 9 for quantization noise and additive noise respectively for different values of SNR. As expected [10], the distances are widely fluctuating over the utterance and they are larger for lower SQNR or SANR. In general, the distances for additive white noise are larger than for quantization. But, as discussed earlier, D_1 does not reflect the true differences in the spectral envelopes, and therefore any conclusions based on it such as the effect of the nature of distortion or bit rate may not be appropriate. The distance (D_2) contours, obtained after modifying the autocorrelation coefficients of the original data, are shown in Figs. 10 and 11. It is very interesting to note that the large distances in Fig. 8 for ADPCM data are reduced to very low values even for the lowest bit rate (20 kBits/sec). The values are uniformly low for all frames and bit rates indicating that ADPCM coding produces negligible changes in the spectral shape. On the other hand the modified distances for

additive noise, although significantly reduced from those in Fig. 9, are consistently much larger than for ADPCM data. Moreover, the values of D_2 are not very much reduced even when the SANR is increased from 10 dB to 27dB.

The effect of modification of autocorrelation coefficients on the spectral shape is illustrated in Figs. 12 to 15 for two frames (5th and 11th) for 2 bit ADPCM and for 10 dB additive white noise distortion. It is clear that the large distances in Figs. 8 and 9 for both types of distortion are due to changes in the dynamic range produced by the distortion. In case of quantization noise the deviation is mainly in the high frequency region, whereas for additive noise the deviation is present throughout the spectrum. After the modification, the original spectrum is brought to the same level as the distorted one in ADPCM case and hence the distance D_2 is very small. In case of additive noise, even after the modification, there exists significant differences between the spectra especially in the bandwidths of formants. The spectra for the 5th frame (Fig. 13) shows an interesting point that the best comparison with distorted spectrum is by a near horizontal line obtained after modification of the original data. This brings out the important differences between the two types of distortion under consideration. The quantization noise affects mainly in the low SNR regions of the spectrum whereas the additive noise affects the entire spectrum. The

formant peaks are relatively unaffected in ADPCM data, except probably in the high frequency region. On the other hand, for additive noise case, the formant peaks are altered both in location and width and also several spurious peaks appear in the smoothed spectrum. For some frames in Fig. 9 there appears to be large distance between ADPCM and original data even after modification. The reason for this is due to large deviation of the modified spectrum in the high frequency region as shown in Fig. 16 for the 11th frame of 3 bit ADPCM. Such occurrences are few and isolated, and one way to overcome the problem is to demphasize the high frequency region in the distance metric.

The total distance for the whole sentence for different cases in Figs. 9 to 12 are computed and plotted in Fig. 17. Although the total distances are significantly reduced in both cases of distortion after modification, the reduction is more striking for ADPCM data. The reduced distances are an indication of the true differences in the spectral envelopes. The somewhat large values of reduced distances for 3 bit and 4 bit ADPCM are due to the isolated peaks in Fig. 9 as explained earlier. It is again evident from Fig. 17 that ADPCM coding has smaller effect on the spectral envelope even for the lowest bit rate whereas additive white noise has significant effect on the spectral envelope even for an SANR of 27 dB. Thus the general conclusion reached by Sambur and Jayant, [10] that the white noise degradation is more severe than quantization noise degradation, is

still valid, although the distance measure (D_1) used by them does not bring out the true differences in the nature of distortion. The measure D_1 is not sensitive to important spectral properties of speech sounds such as formant peaks and their bandwidths, as reported in [10]. The modified measure (D_2) is more sensitive to changes in spectral envelope and should be considered for comparing spectra. The same general conclusions on the effect of these two types of distortion are also reported by Gibson [14] from a theoretical study of LP analysis of distorted speech.

In order to verify that the true differences in spectral envelopes are retained in the distance measure even after the modification of the autocorrelation coefficients, we computed the distances D_1 and D_2 between adjacent frames for the original and distorted speech data as shown in Figs. 18 to 21. Large differences in the spectral envelopes between adjacent frames are reflected in the measure D_2 as well, although its values are a little smaller, indicating that D_2 is a useful measure of spectral differences. There is however a possibility that the dynamic range is not properly compensated in some cases because always only the first autocorrelation set is modified relative to the second. It is more appropriate to use the two way modification suggested in Sec. III,

since the spectra being compared here are for different sounds. When this is done the D_3 contours obtained are as shown in Figs. 22 and 23 for the two types of distortion. Most of the spectral differences are preserved as before. The lower absolute values for low bit ADPCM compared to 5 bit ADPCM or original data indicate the effect of spectral dynamic range even after modification.

Normalized linear prediction error (η_M) is another parameter which is used to study the behavior of spectral shape [2]. It is defined as

$$\eta_M = 1 + \sum_{k=1}^{M} a(k)\, r(k) \qquad (23)$$

where {a(k)} and {r(k)} are the LPCs and the normalized autocorrelation coefficients respectively. Figs. 24 and 25 show the plots of η_M contour for the distorted and original data. For ADPCM data the η_M contour has the same shape as that for the original even after modification. For additive noise, on the other hand, the η_M contours are very much different from the one for original data even for the case of SANR=22.5 dB. Moreover the modification of the autocorrelation coefficients of the original data increased the error as expected, and the modified values are nearly one for several frames of data. This is due to the fact that the modified spectrum is nearly flat, i.e., all other autocorrelation coefficients are negligible compared to R(0). Despite this the distances are reduced as shown in Fig. 10, because the large distances (D_1) in Fig. 9 are not a measure of differences in spectral envelope but are mainly a measure of differences in spectral dynamic range.

Comparison of absolute values of the normalized error may not always give an indication of differences in spectral envelopes either, as can be seen from the expression for η_M in terms of the zeroth coefficient $\hat{c}(0)$ of cepstrum and the signal energy R(0) [2].

$$\eta_M = \exp[\hat{c}(0)] \; / \; R(0) \qquad (24)$$

where

$$\hat{c}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \left[\sum_{k=0}^{M} R(k) \cos(\omega k) \right] d\omega \qquad (25)$$

Values of η_M can be altered by R(0) even though the spectral envelope determined by R(1), R(2), ... R(M) remain unchanged.

VI. CONCLUSIONS

We have shown that comparison of spectra by distance measures based on LP smoothing do not yield the true differences in the spectral envelopes. This is because the LPCs are altered significantly by changes in the zeroth autocorrelation coefficient alone which does not carry the spectral envelope information. The large values of distances obtained in studies using the LP distance measures are mostly due to differences in the dynamic range of the spectrum. We have demonstrated this fact by showing that the large values of LP distance between the original and distorted data are reduced to very small values when the spectra are brought to a common level of dynamic range by altering R(0) alone. Quantative assessment of degradation using the modified distance measure show that the

spectral envelope is much less susceptible to quantization noise distortion than to additive noise distortion. The same general conclusions were obtained by Gibson [14] through theoretical studies and by Sambur and Jayant [10] through experimental studies of LP analysis of distorted speech. However, the threshold levels of significant differences used in [10] are not valid for the new measure proposed in this paper. The main result of our study is that ADPCM coding does not affect the spectral envelope significantly even for the lowest bit rate case. In contrast, even small quantity of additive white noise seems to effect the spectral shape. These results are valid from a perceptual angle also. It has been observed [3] that there is little correlation between distance measures and perceptual changes in formants which is obvious from the results reported in this paper.

In several applications such as voiced/unvoiced/silence classification [15], varible frame rate vocoding [16] etc., the distance measure D_2 may be more appropriate than D_1 which is currently being used. The conclusion on ADPCM data explain to some extent the high recognition scores obtained for the distorted data in a speech recognition experiment [18].

ACKNOWLEDGEMENTS

The authors wish to thank Dr. N. S. Jayant of Bell Laboratories for many useful suggestions throughout this work.

REFERENCES

[1] D. R. Reddy,ed., Speech Recognition: Invited Papers of the IEEE Symposium, New York, Academic Press, 1975.

[2] J. Makhoul, " Linear Prediction: A tutorial Review," Proc. IEEE, vol.63, 561-580,1975.

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

[4] L. L. Pfeifer, "Inverse Filter for Speaker Identification," SCRL, Santa Barbara, CA, Final Report RADCTR-74-214, 1974.

[5] J. D. Markel and A. H. Gray, "Linear Prediction of Speech, New York , Springer-Verlag, 1976.

[6] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., vol. 55, 1304-1312, 1974.

[7] D. T. Magill, "Adaptive Speech Compression for Packet Communication System," in Conf. Rec., IEEE Telecommunications Conf. 1973

[8] F. Itakura, "Minmum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, 67-72, 1975.

[9] R. W. Schafer and L.R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer., vol. 47, 634-648, 1970.

[10] M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantization Noise or Additive White Noise," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, 488-494, 1976.

[11] A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, vol ASSP-23, 169-175, 1975.

[12] J. Makhoul, R. Viswanathan, L. Cosell and W. Russel, Natural Communication with Computers: Speech Compression Research at BBN, BBN Rept. No. 2976II, Bolt, Beranek and Newman Inc, Cambridge, MA., 1974.

[13] P. V. de Souza, "Statistical Tests and Distance Measures for LPC Coefficients," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, 554-558, 1977.

[14] J. D. Gibson, "Theory of LPC Analysis from Distorted input Speech," TCSL Res. Memo. 77-15, Communication and Control Sys. Lab., Texas A&M Univ., College Station, Texas, 1977.

[15] L. R. Rabiner and M. R. Sambur, "Voiced-Unvoiced-Silence Detection Using the Itakura LPC Distance Measure," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, 338-343, 1977.

[16] B. Yegnanarayana and D. R. Reddy, "Performance of Harpy Speech Recognition System for Speech Input with Quantization Noise," J. Acoust. Soc. Amer., vol. 62, K13(A), 1977.

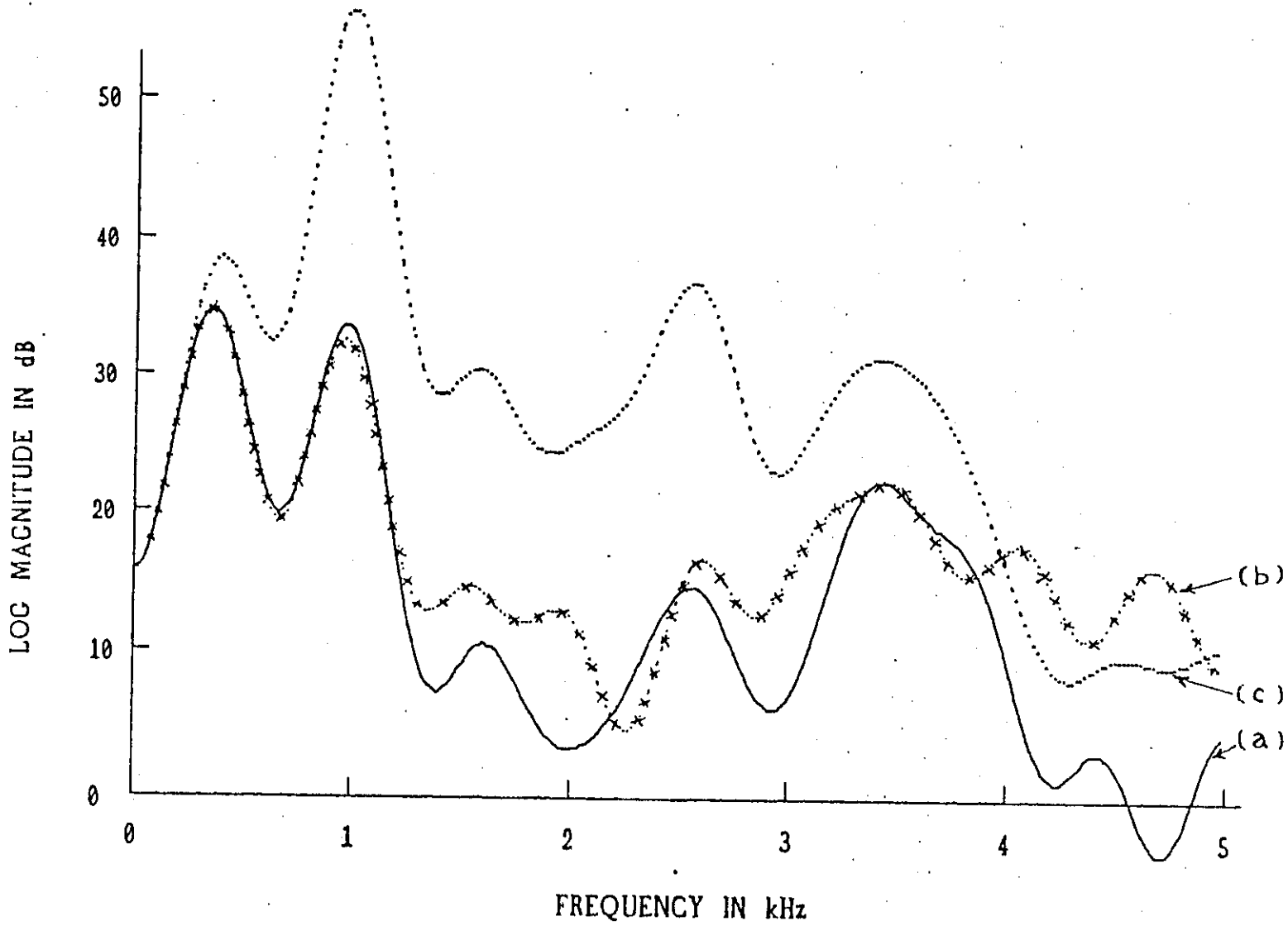

Fig. 1 Example of cepstrally smoothed spectra
(a) undistorted data (b) additive spectral distortion
(c) multiplicative spectral distotion

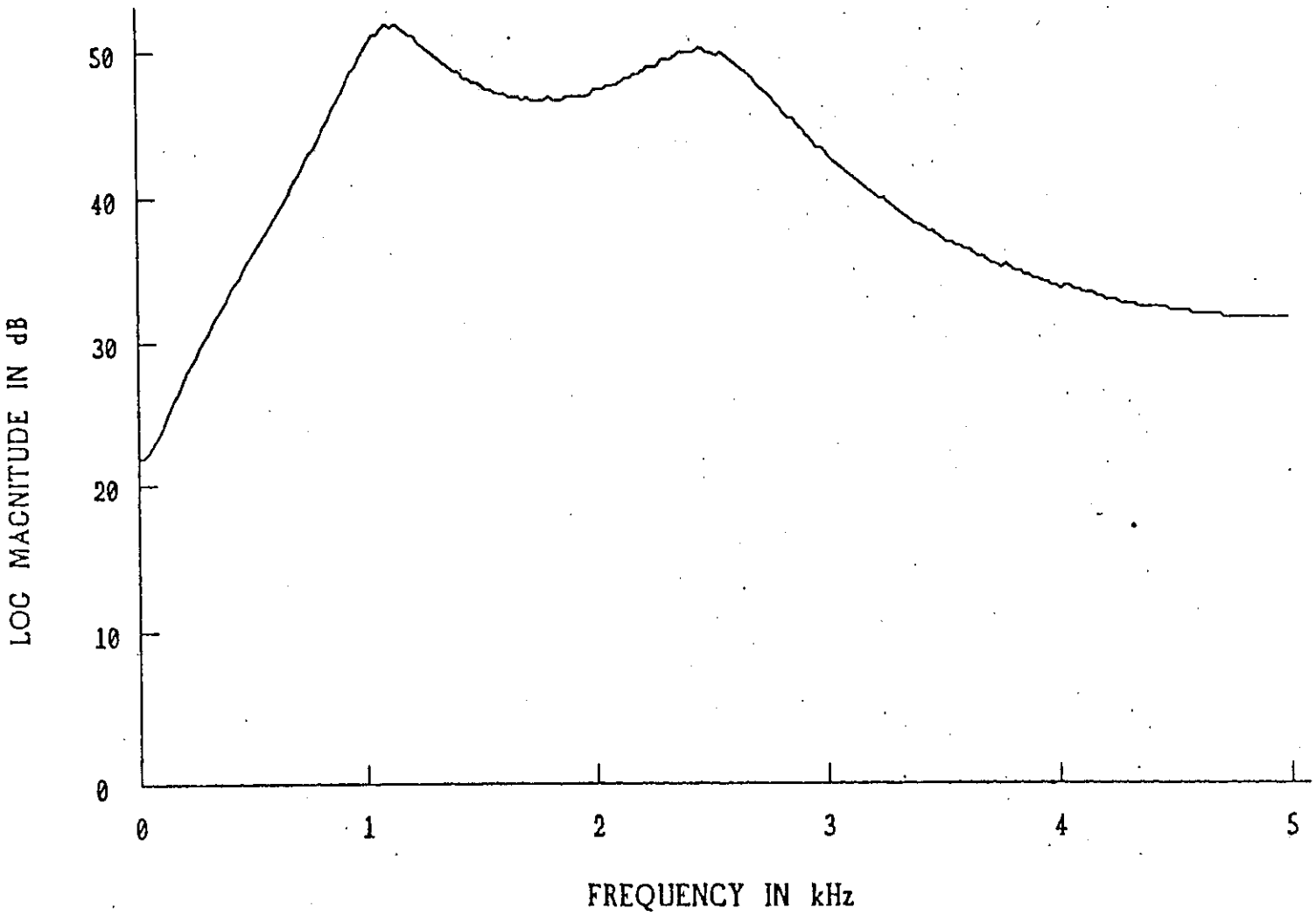

Fig. 2 Frequency responce of multiplicative spectral distotion used in Fig. 1

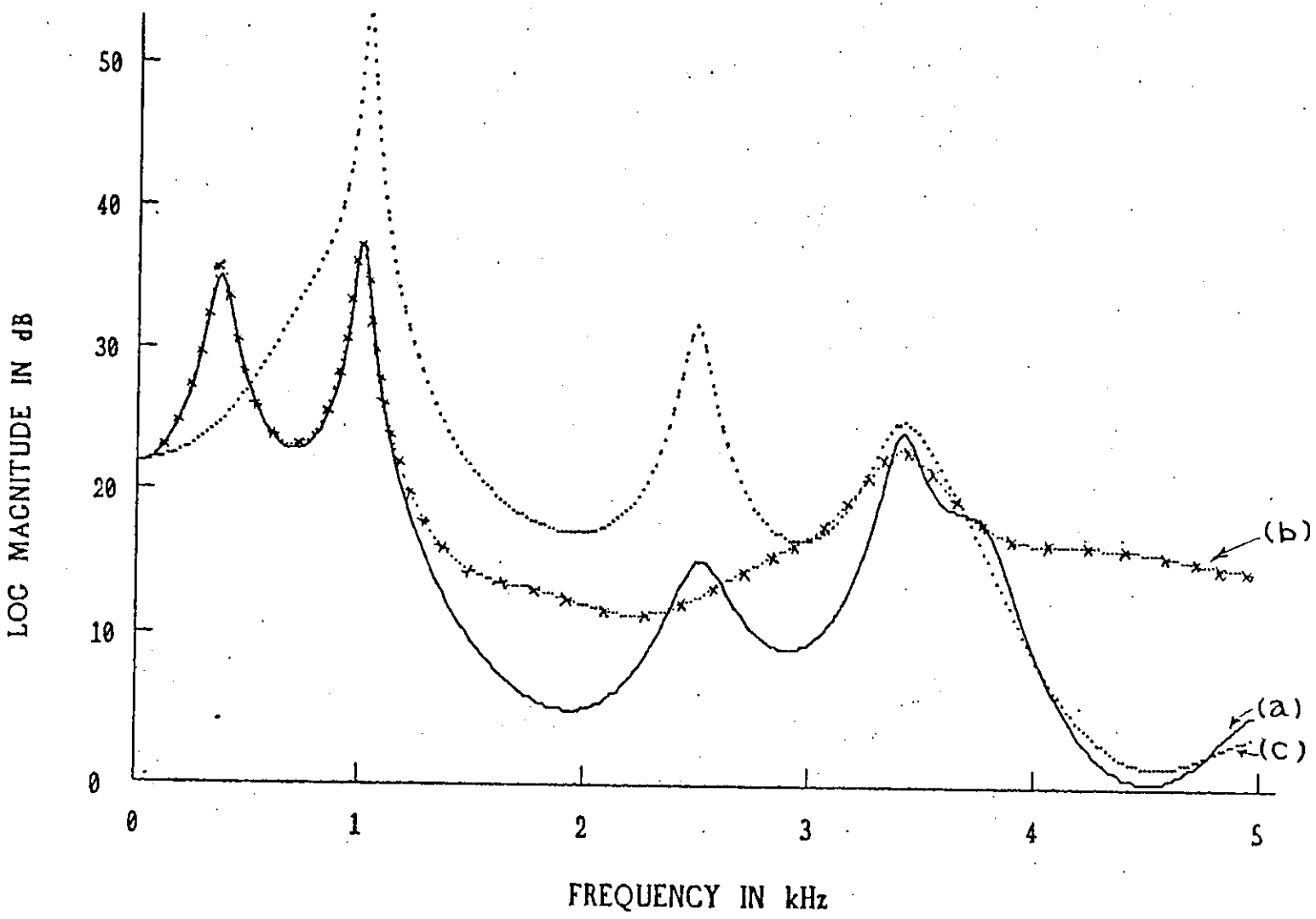

Fig. 3 LP smoothed spectra for the example in Fig. 1

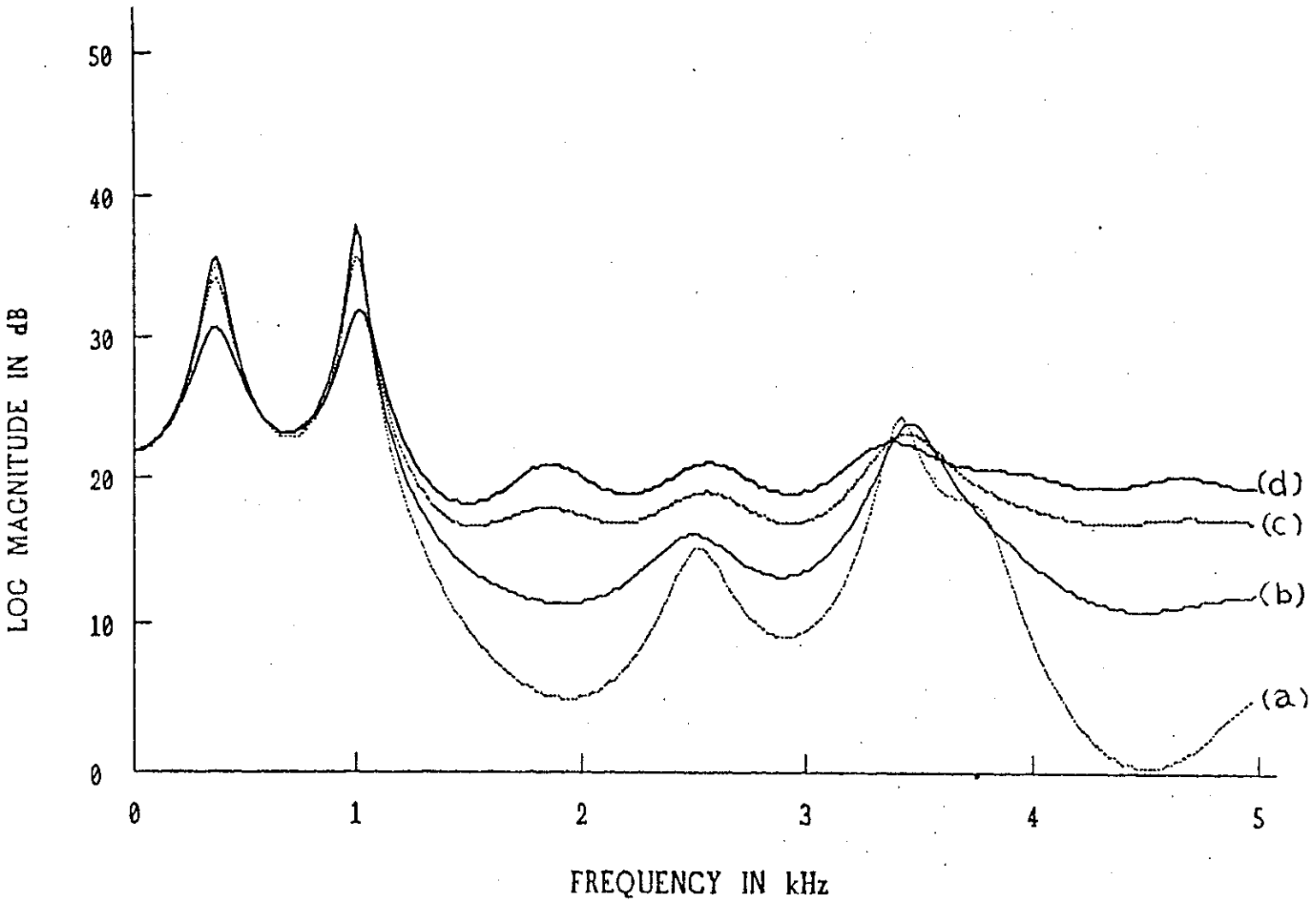

Fig. 4a LP smoothed spectra for different values of R(0) for a voiced segment of speech
(a) 1.00 R(0) (b) 1.05 R(0) (c) 1.20 R(0) (d) 1.45 R(0)

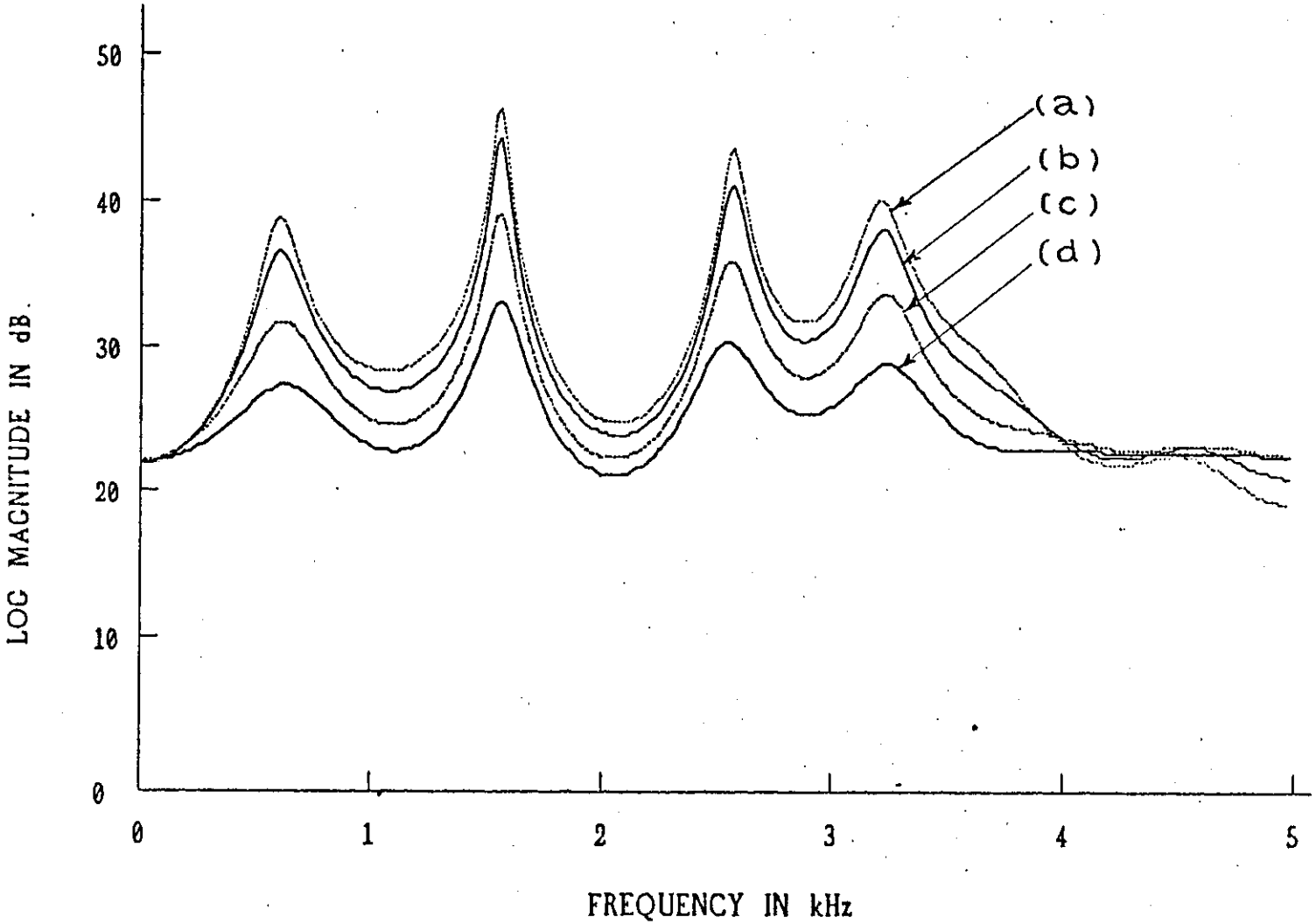

Fig. 4b LP smoothed spectra for different values of R(0) for a unvoiced segment of speech
(a) 1.00 R(0) (b) 1.05 R(0) (c) 1.20 R(0) (d) 1.45 R(0)

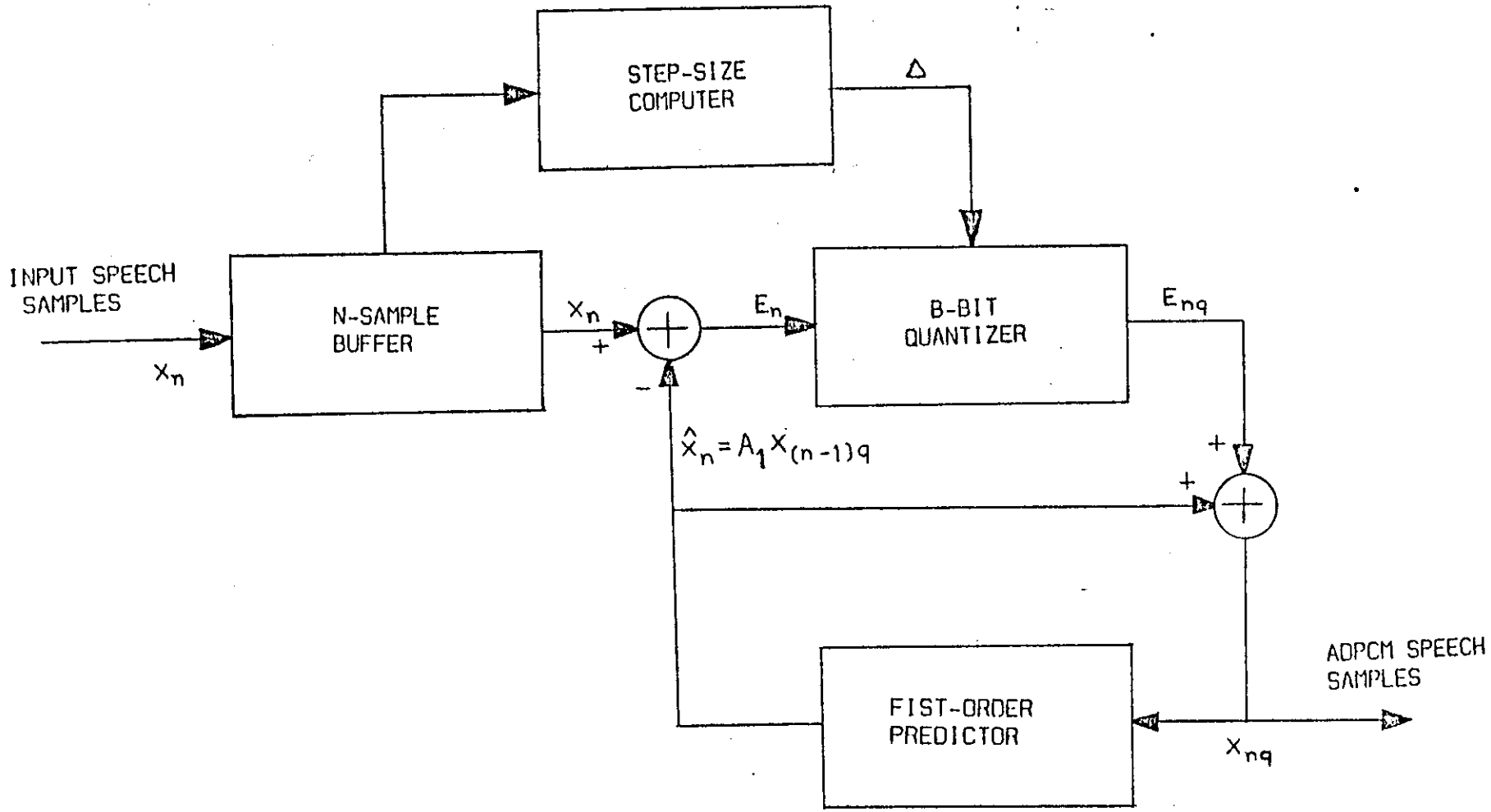

Fig. 5 Block diagram for generating ADCM data

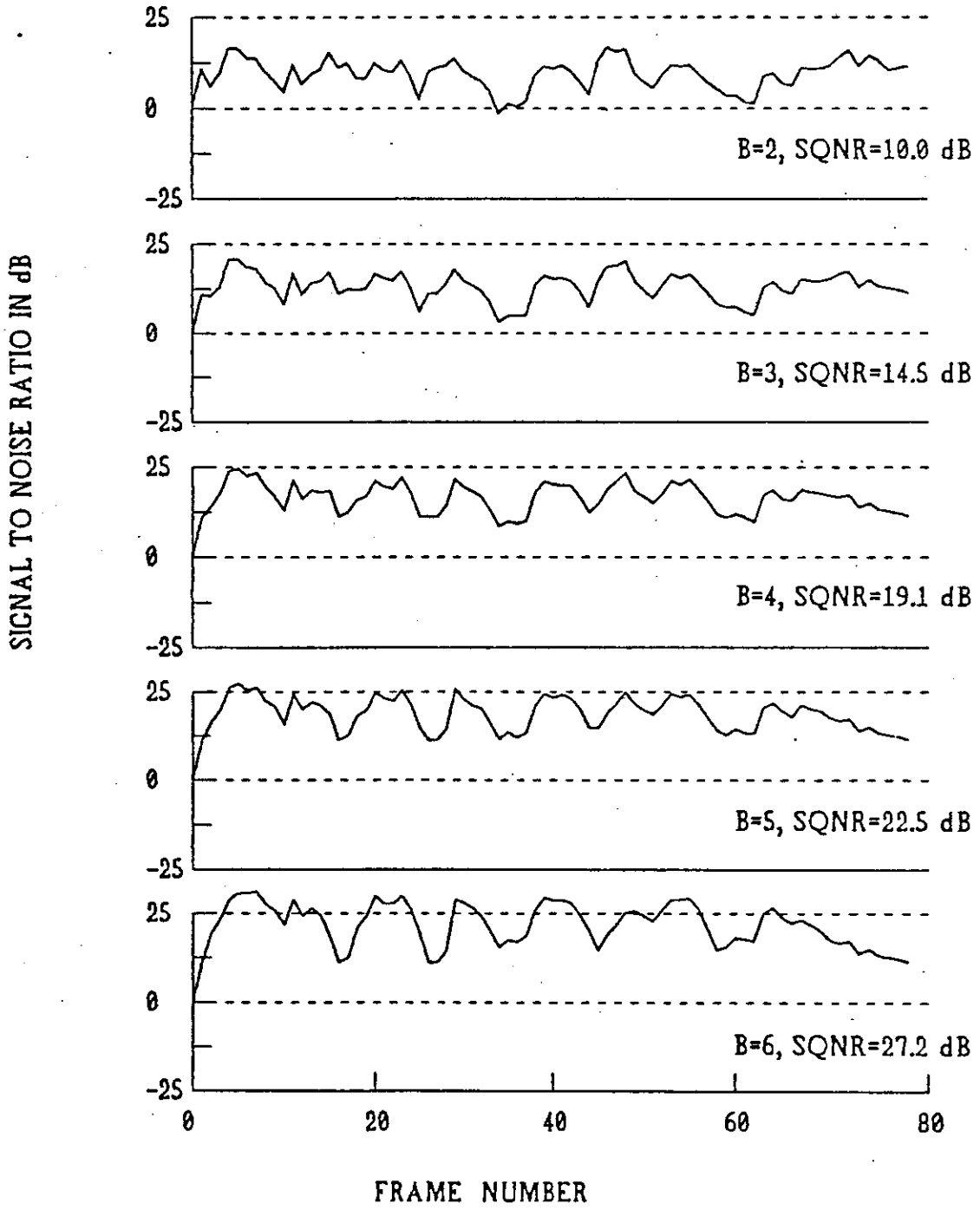

Fig. 6 SNR contours for ADPCM data

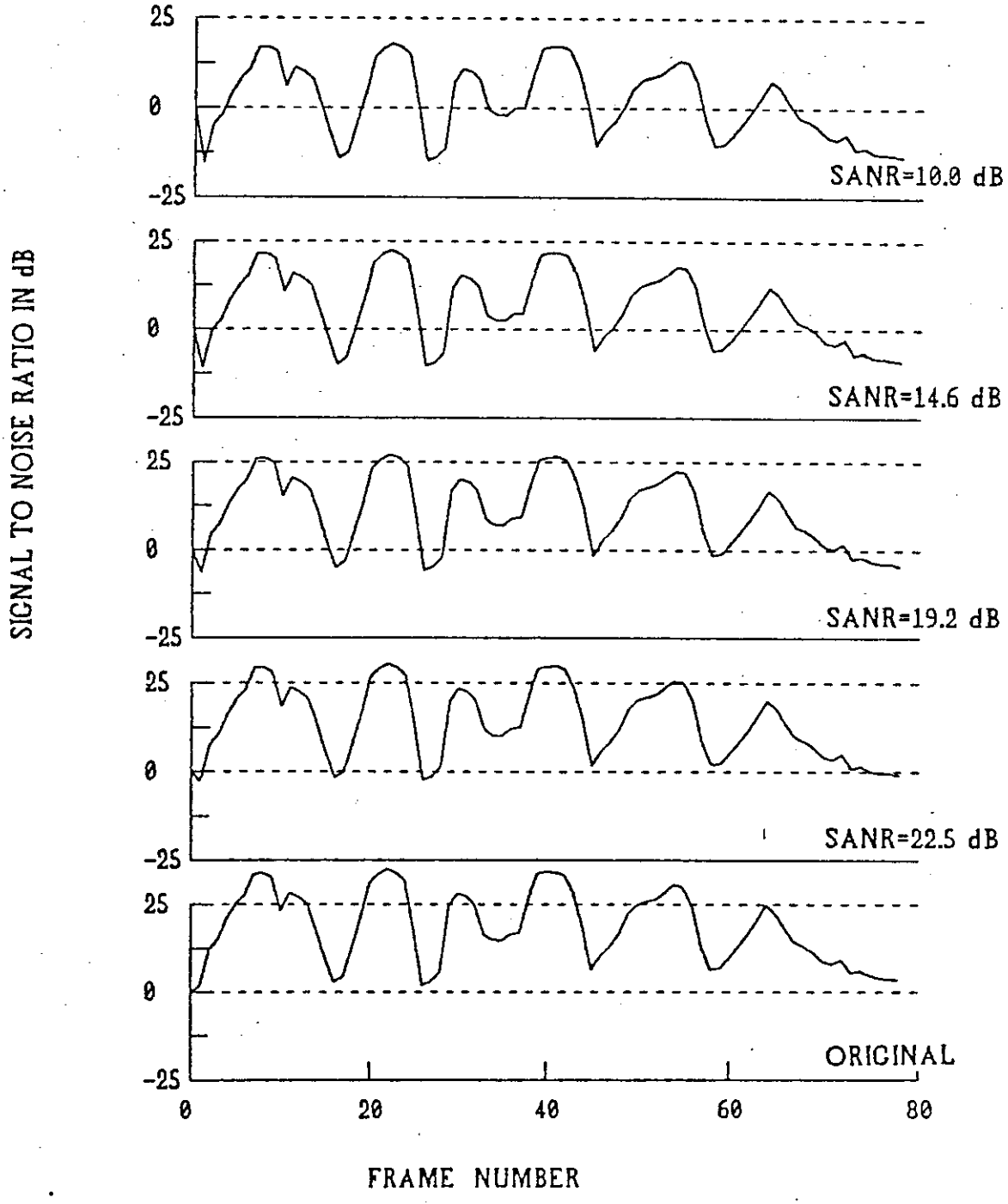

Fig. 7 SNR contours for additive noise distortion

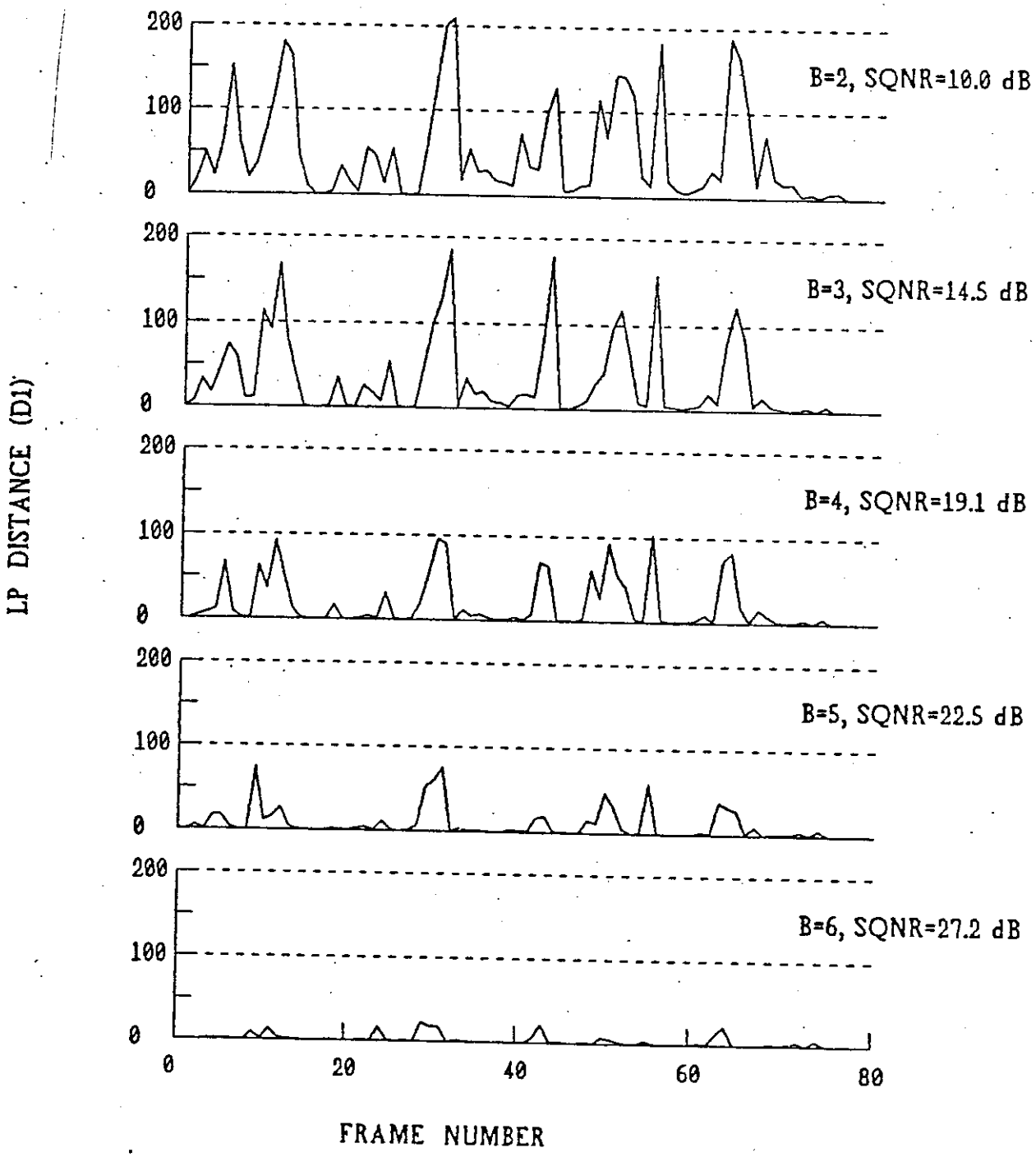

Fig. 8 LP distance (D) contours for ADPCM data
1

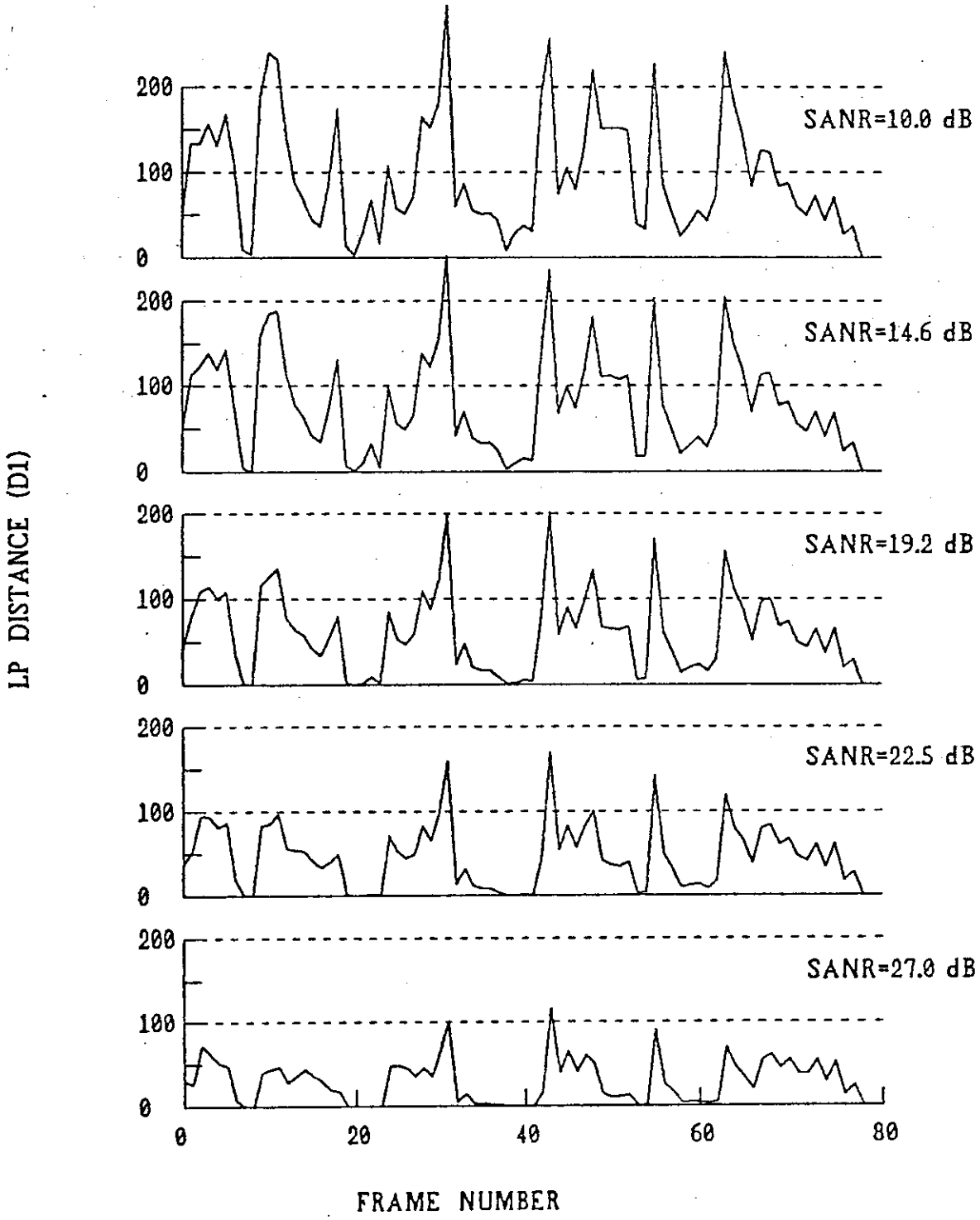

Fig. 9 LP distance (D_1) contours for additive noise distortion

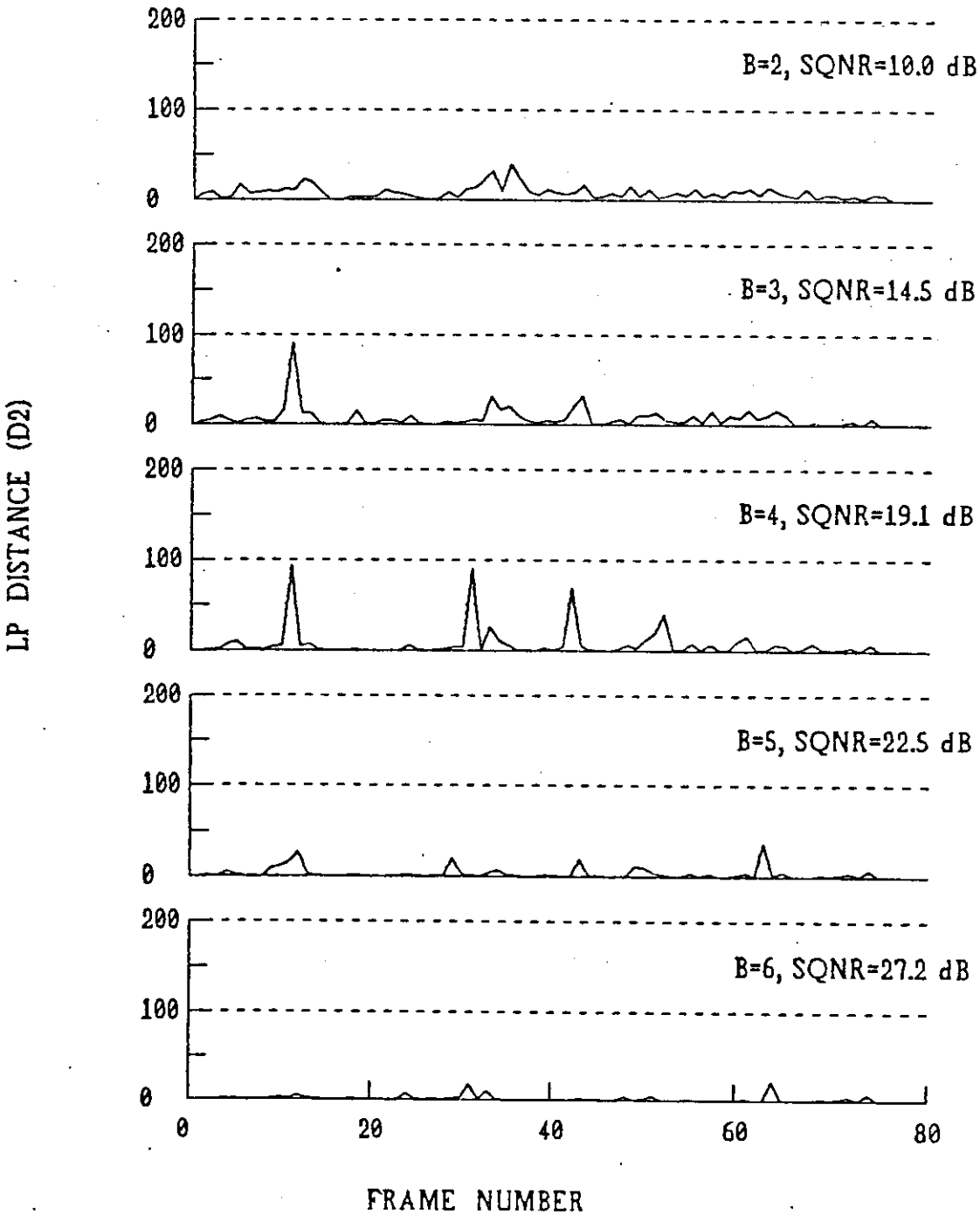

Fig. 18 Modified LP distance (D_2) contours for ADPCM data

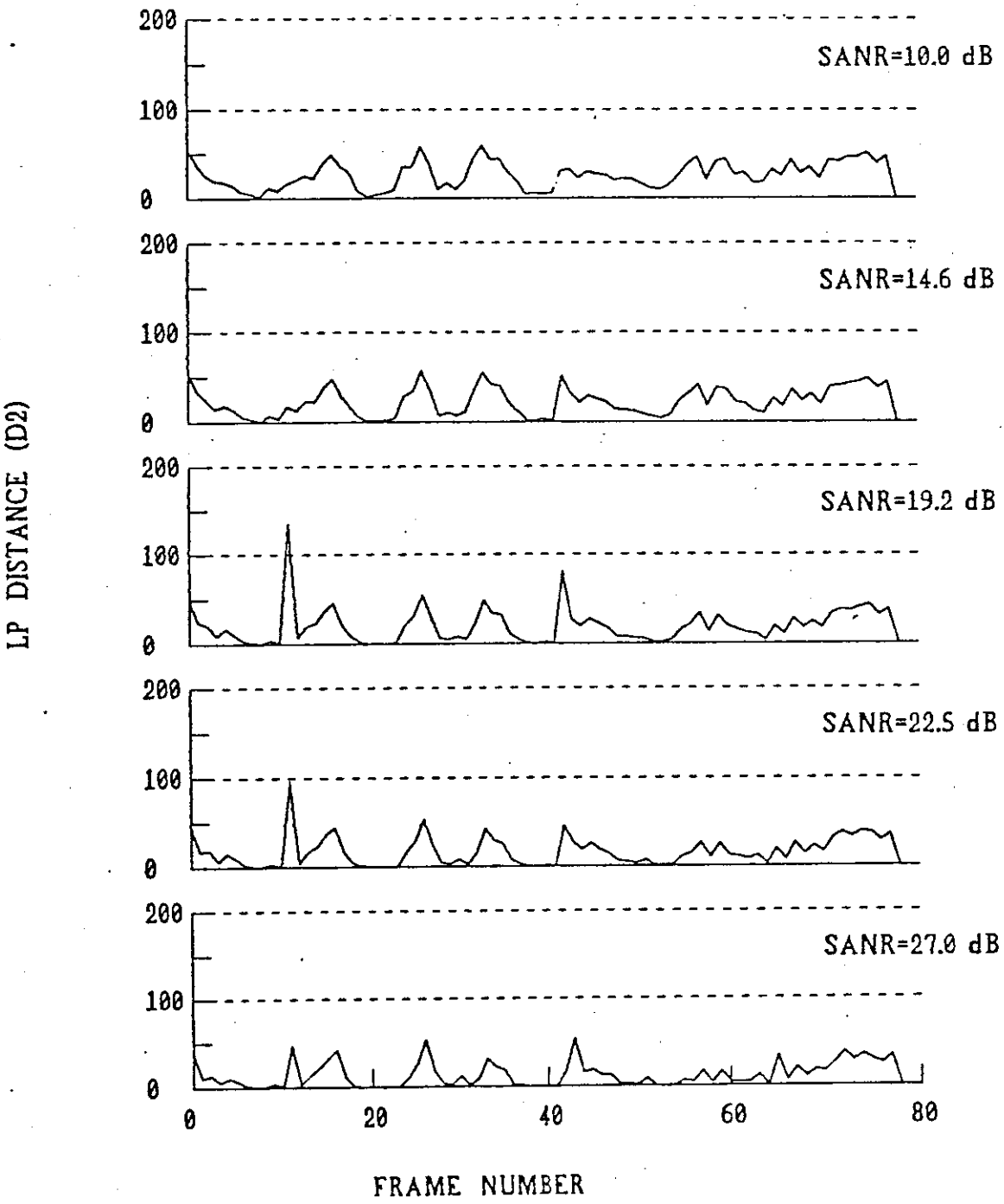

Fig. 11 Modified LP distance (D_2) contours for additive noise distortion

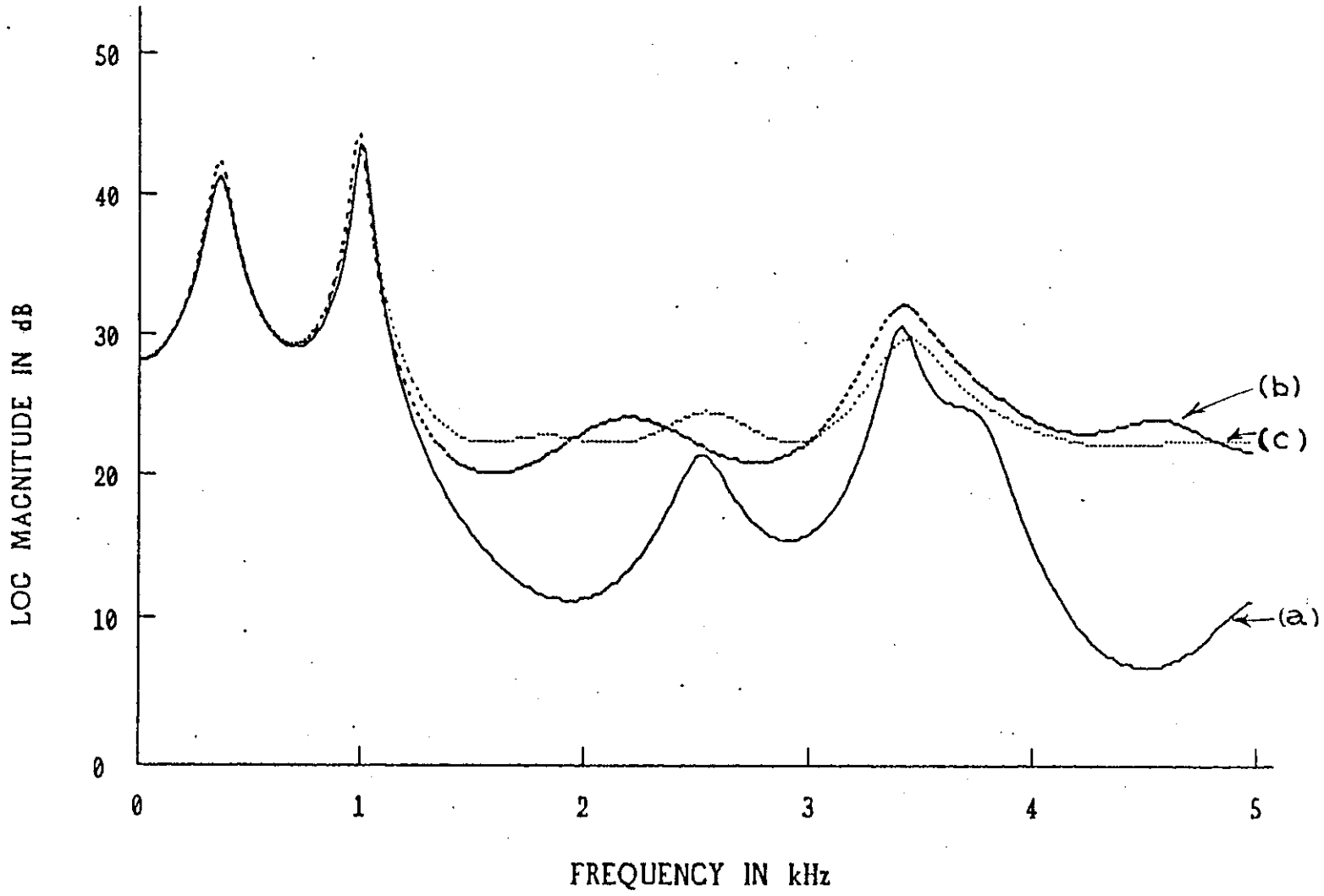

Fig. 12 Effect of proposed modification on LP spectrum of 2 bit ADPCM data - 5th frame
(a) original (b) distorted (c) modified original

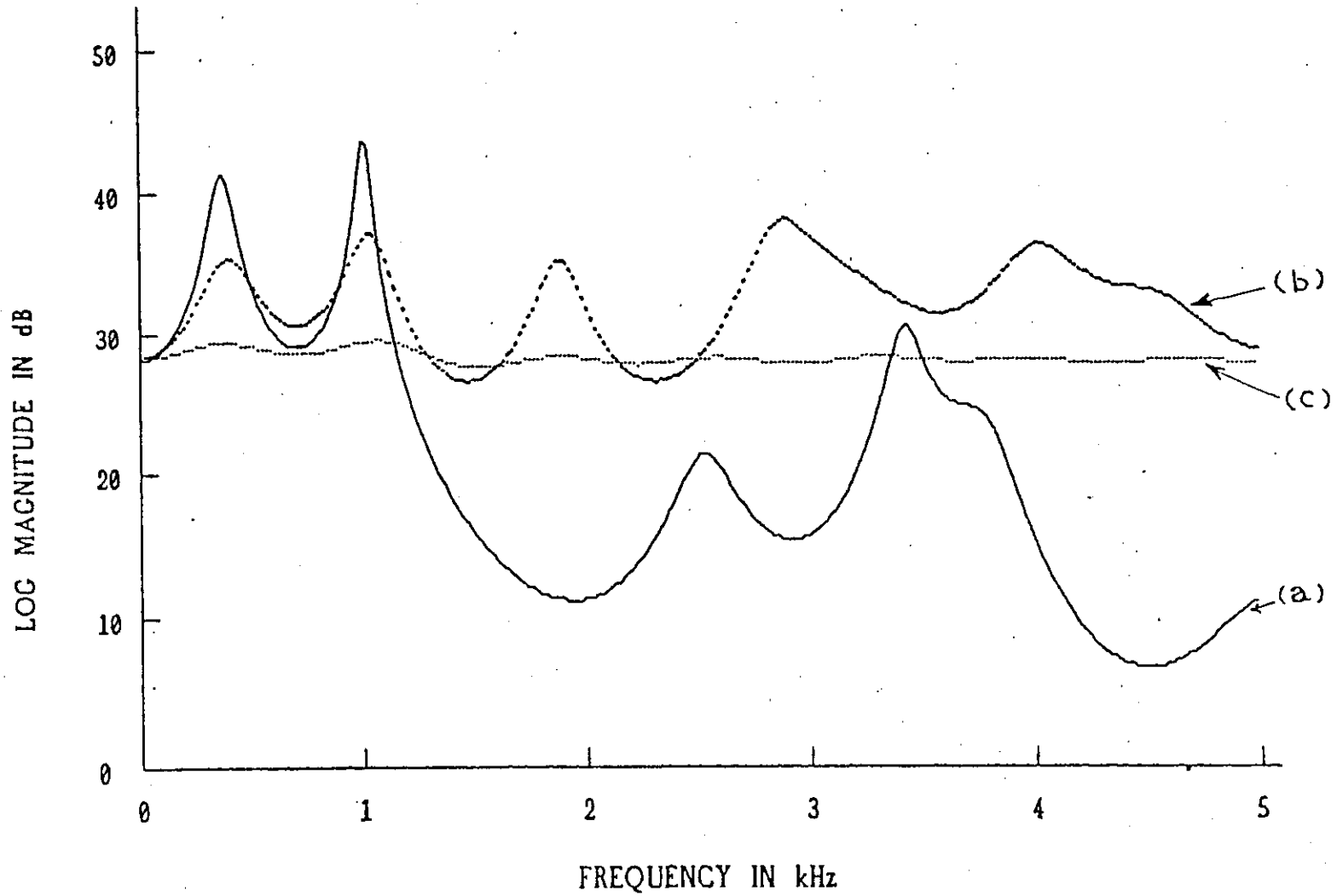

Fig. 13 Effect of proposed modification on LP spectrum of additive noise distortion data (SANR=10 dB) - 5th frame (a) original (b) distortion (c) modified original

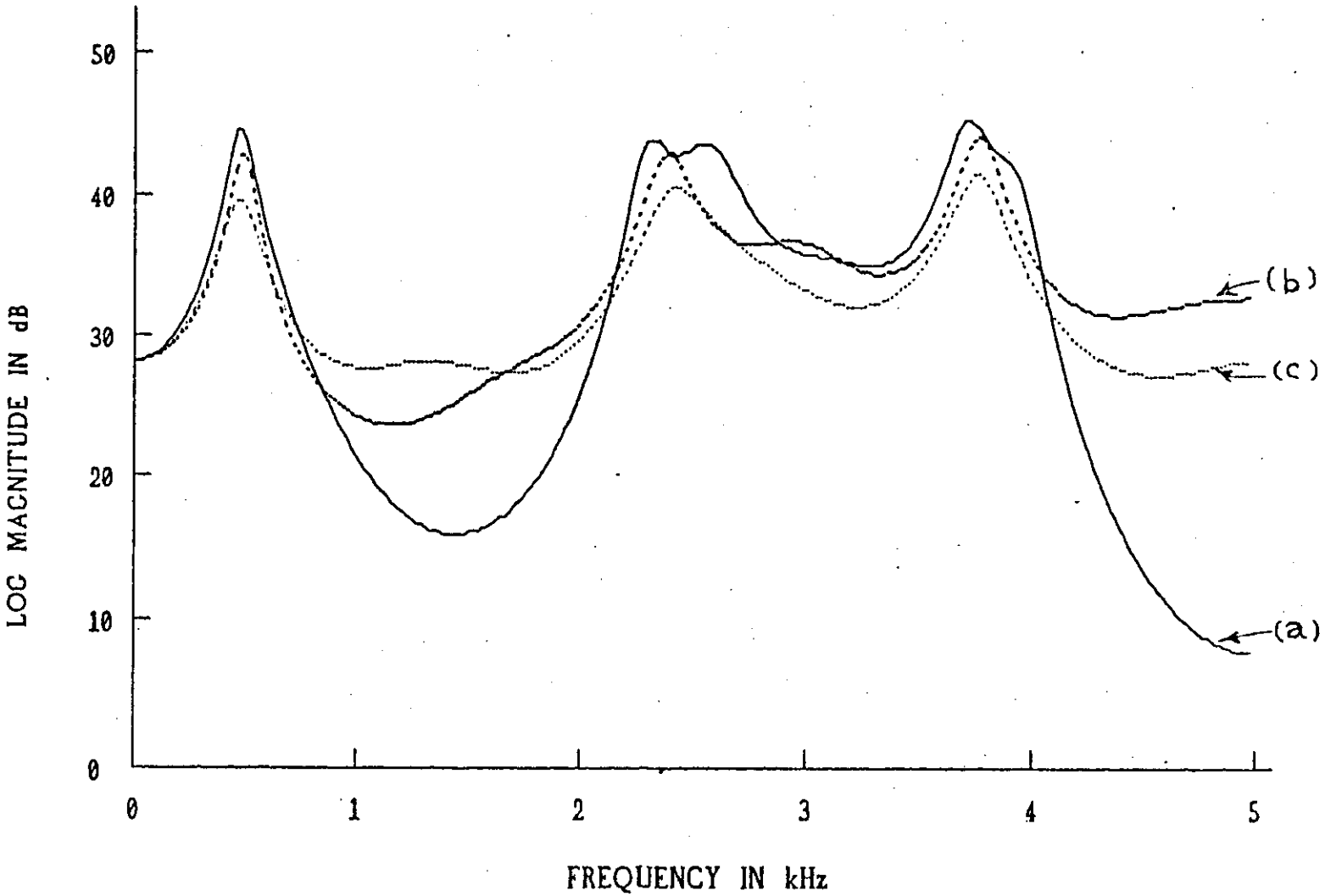

Fig. 14 Effect of proposed modification on LP spectrum of 2 bit ADPCM data - 11th frame
(a) original (b) distortion (c) modified original

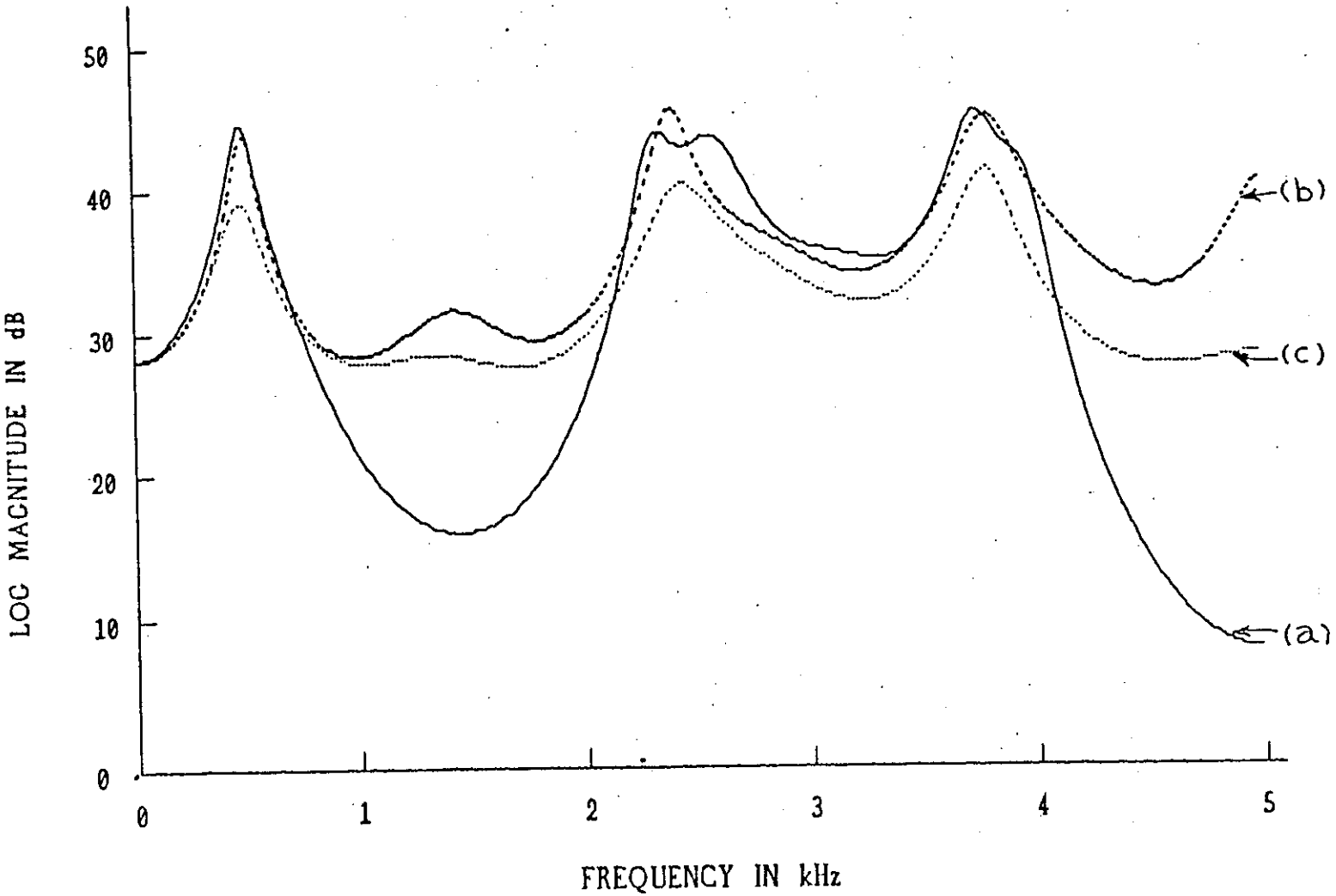

Fig. 15 Effect of proposed modification on LP spectrum of additive noise distortion data (SANR=10 dB) - 11th frame (a) original (b) distorted (c) modified original

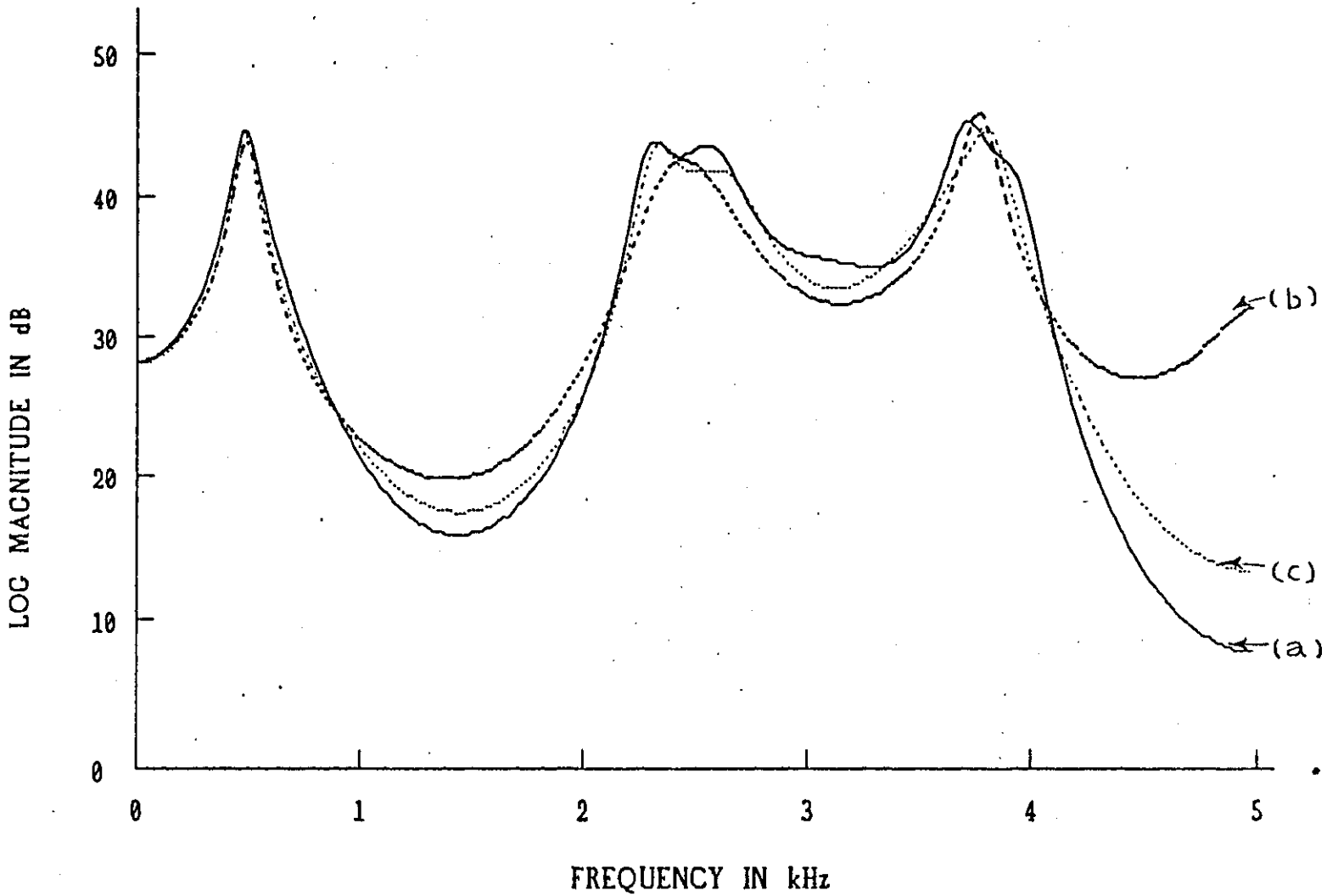

Fig. 16 Effect of proposed modification on LP spectrum of 3 bit ADPCM data - 11th frame
(a) original (b) distorted (c) modified original

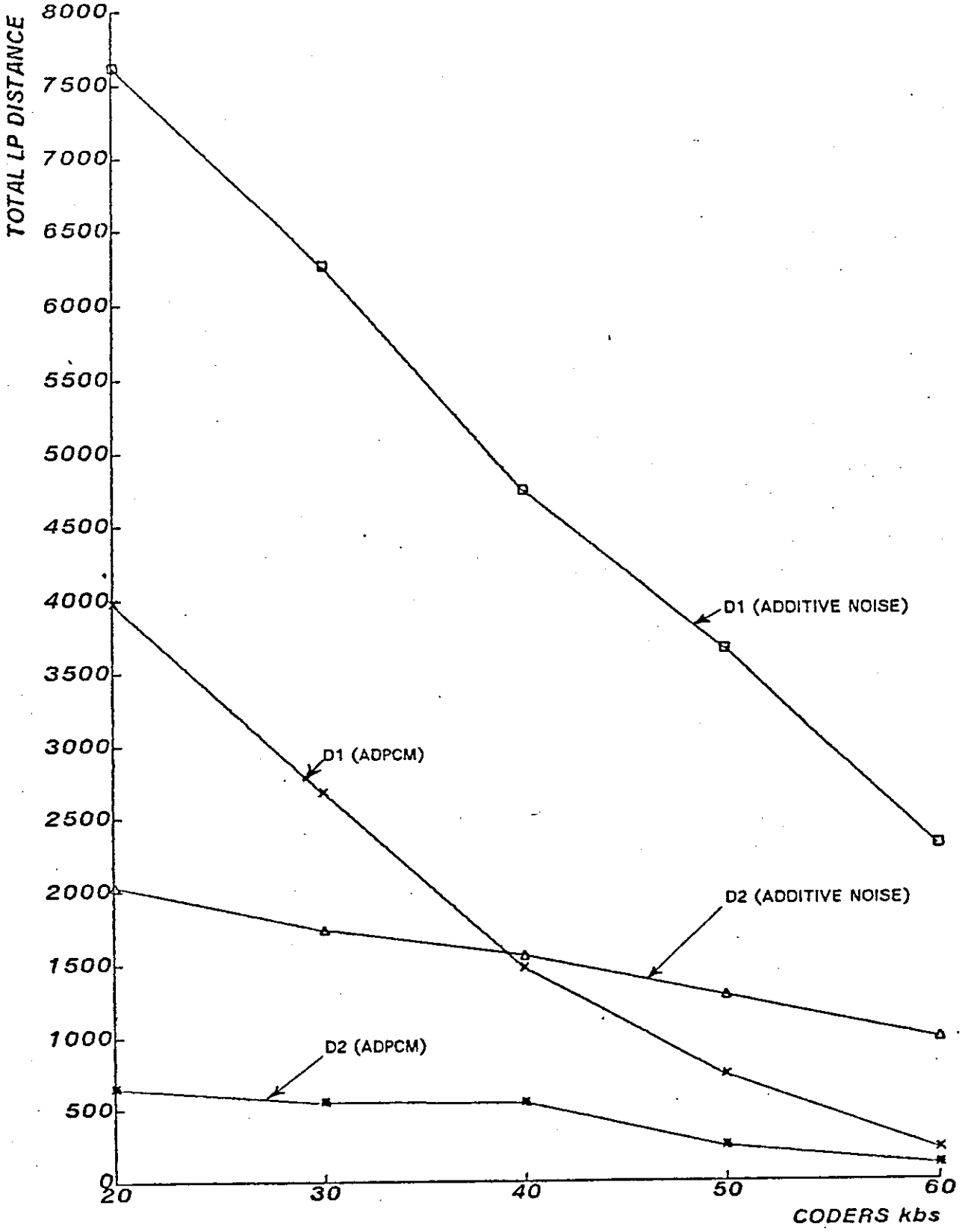

Fig. 17 Total LP distance for different cases of distorted data

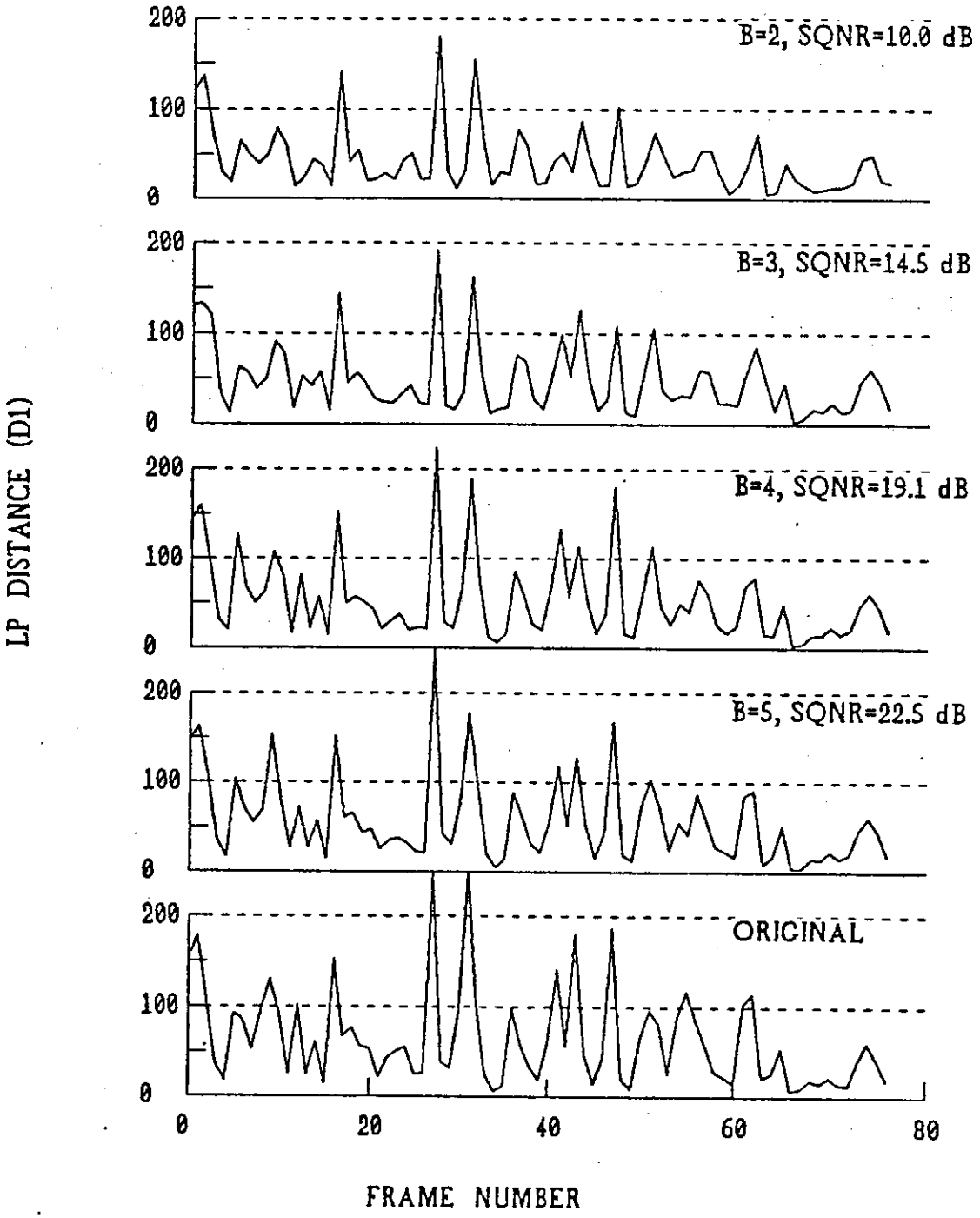

Fig. 18 LP distance (D_1) between adjecent frames for original ADPCM data

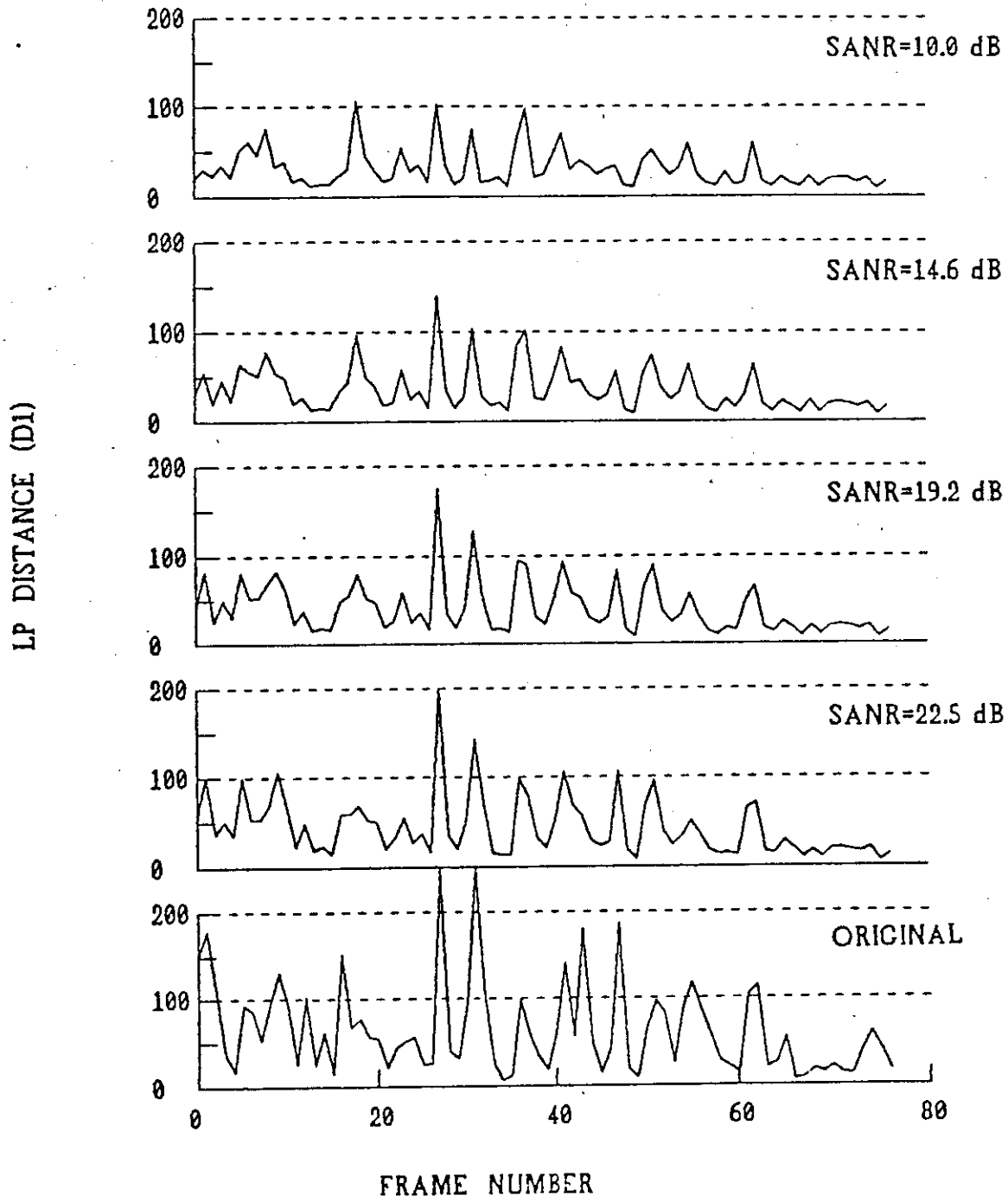

Fig. 19 LP distance (D_1) between adjecent frames for original additive noise distortion data

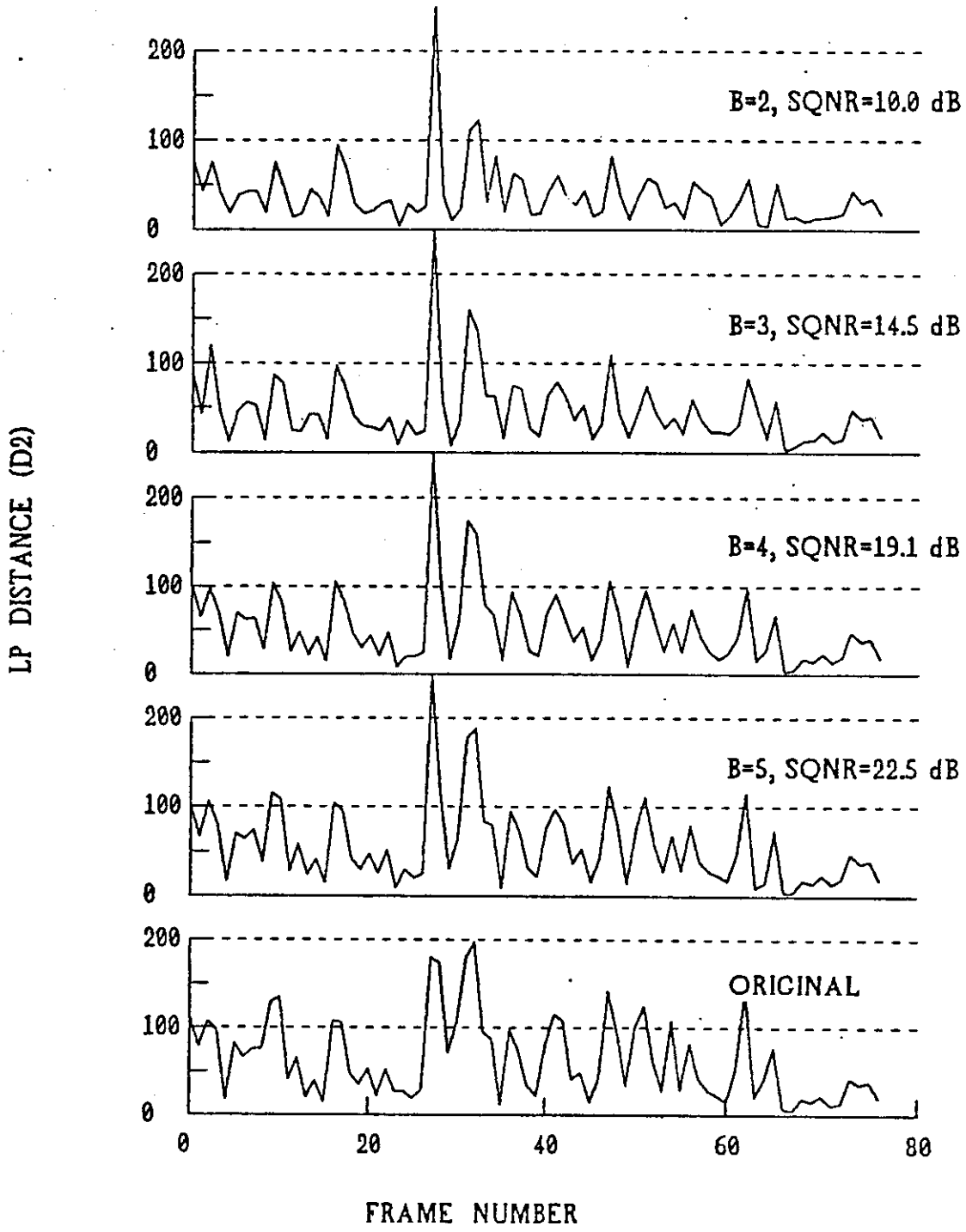

Fig. 20 Modified LP distance (D_2) between adjecent frames for original and ADPCM data

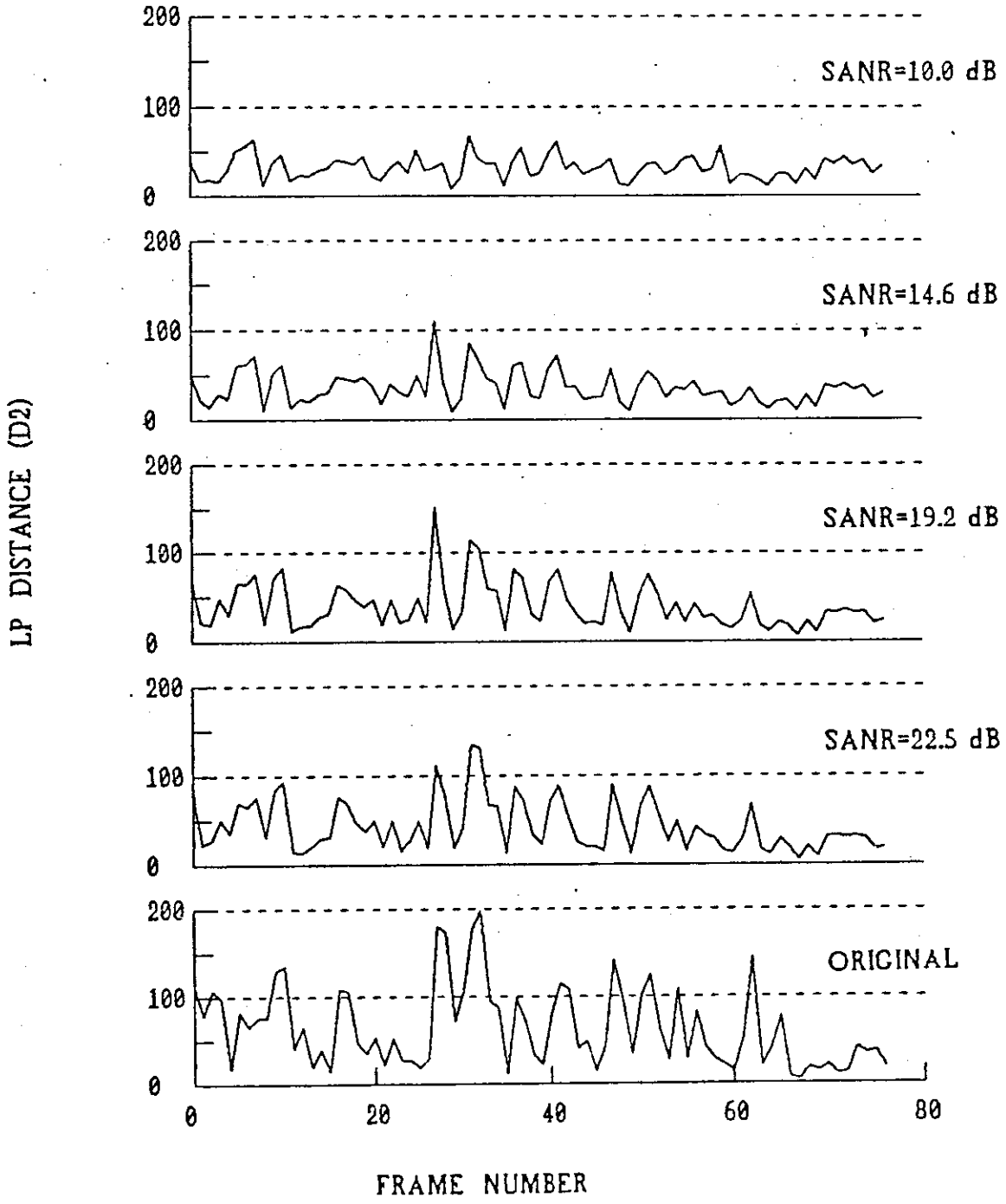

Fig. 21 Modified LP distance (D_2) between adjecent frames for original and additive noise distortion data

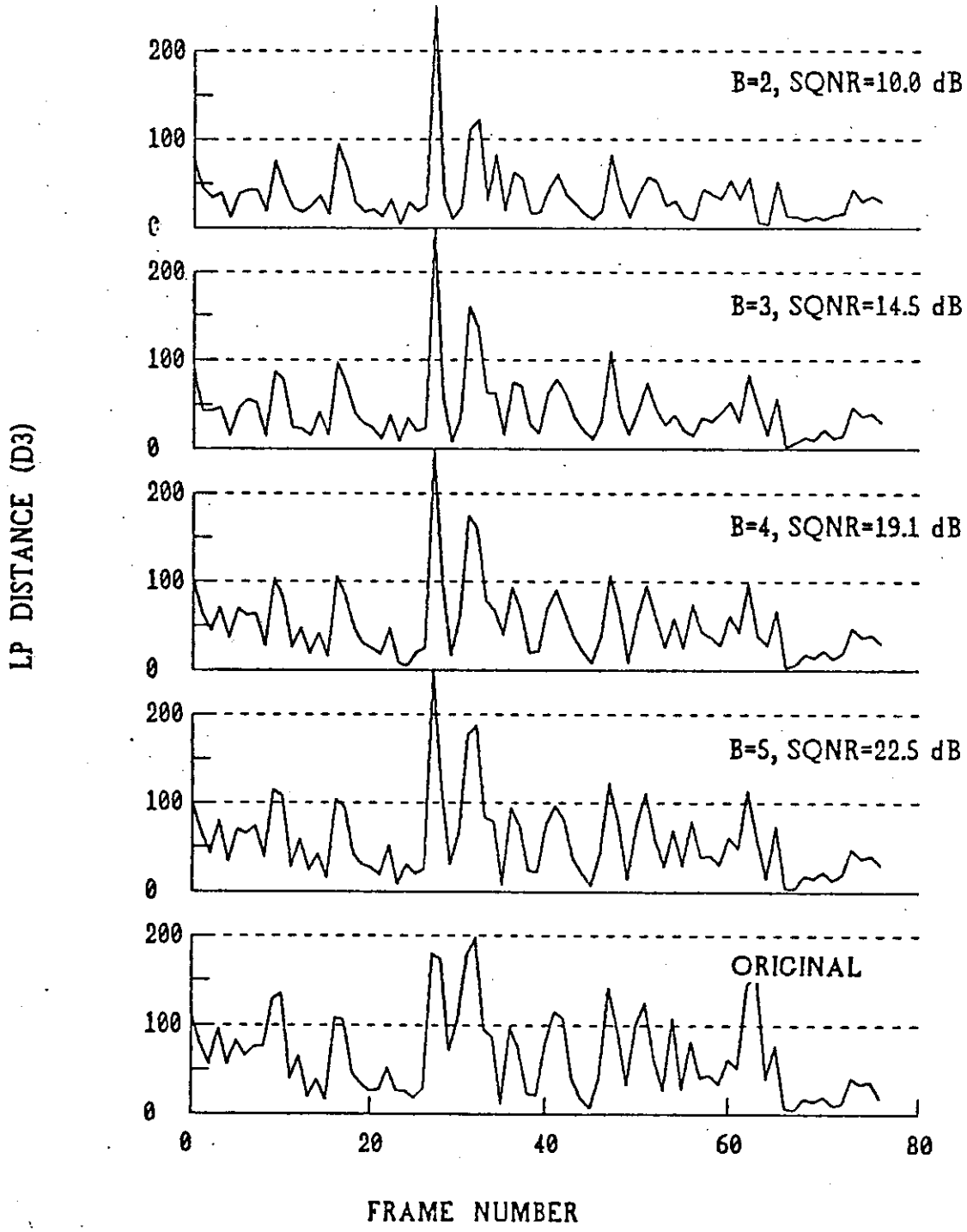

Fig. 22 LP distance (D_3) after two way modification for original and ADPCM data

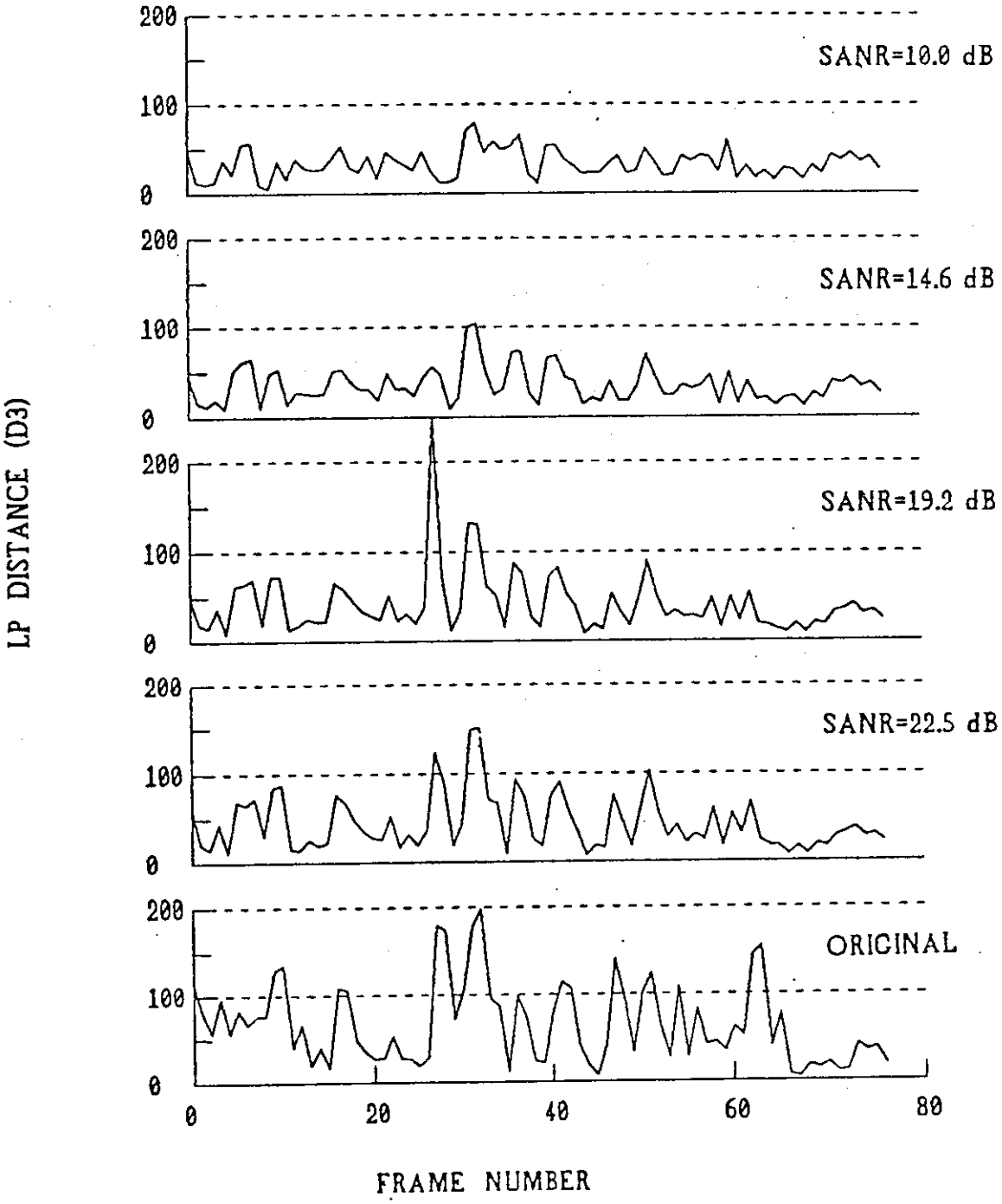

Fig. 23 LP distance (D_3) after two way modification for original and additive noise distortion data

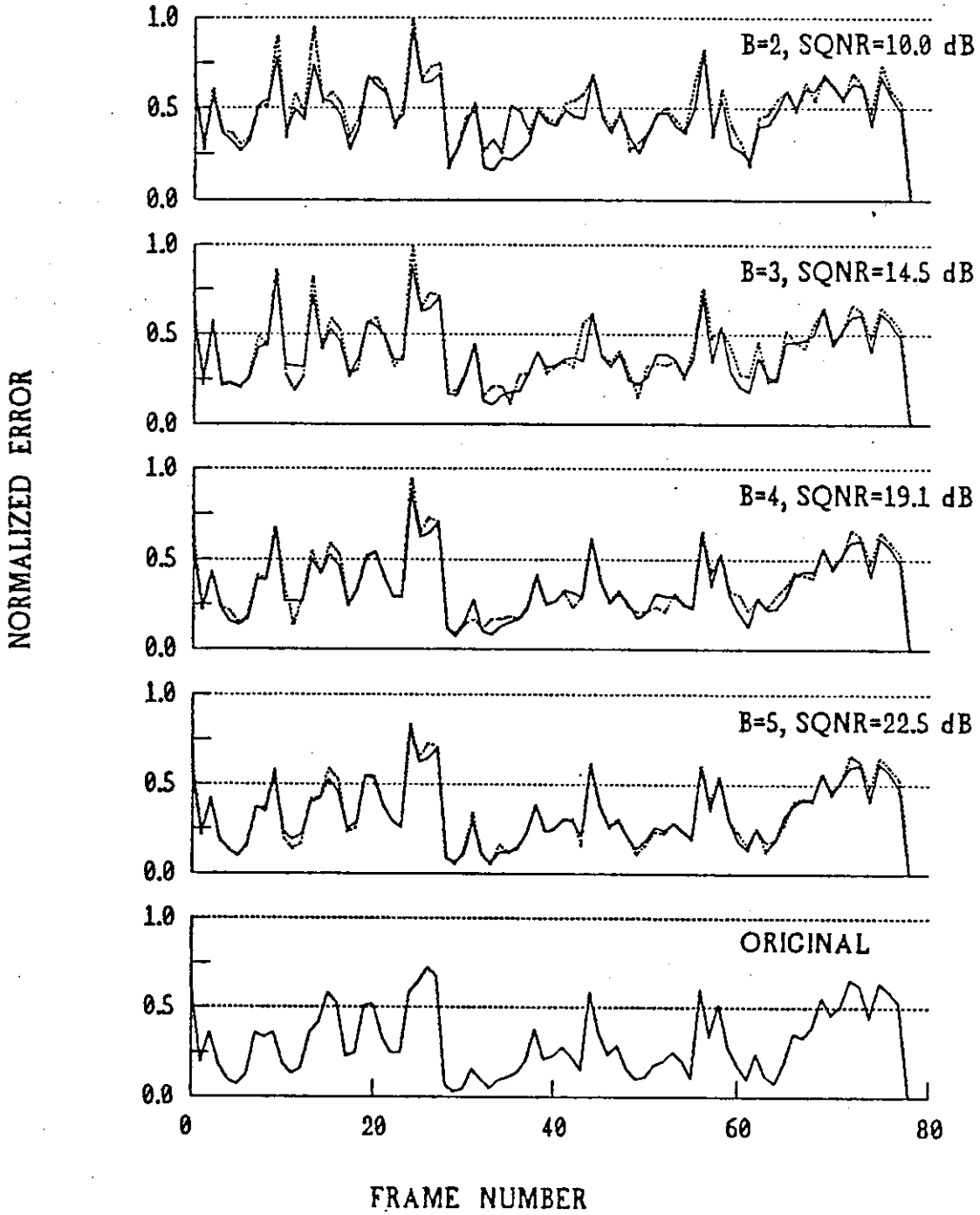

Fig. 24 Normalized LP error (η_{r1}) contour for original data, ADPCM data and modified original data (shown by dotted line)

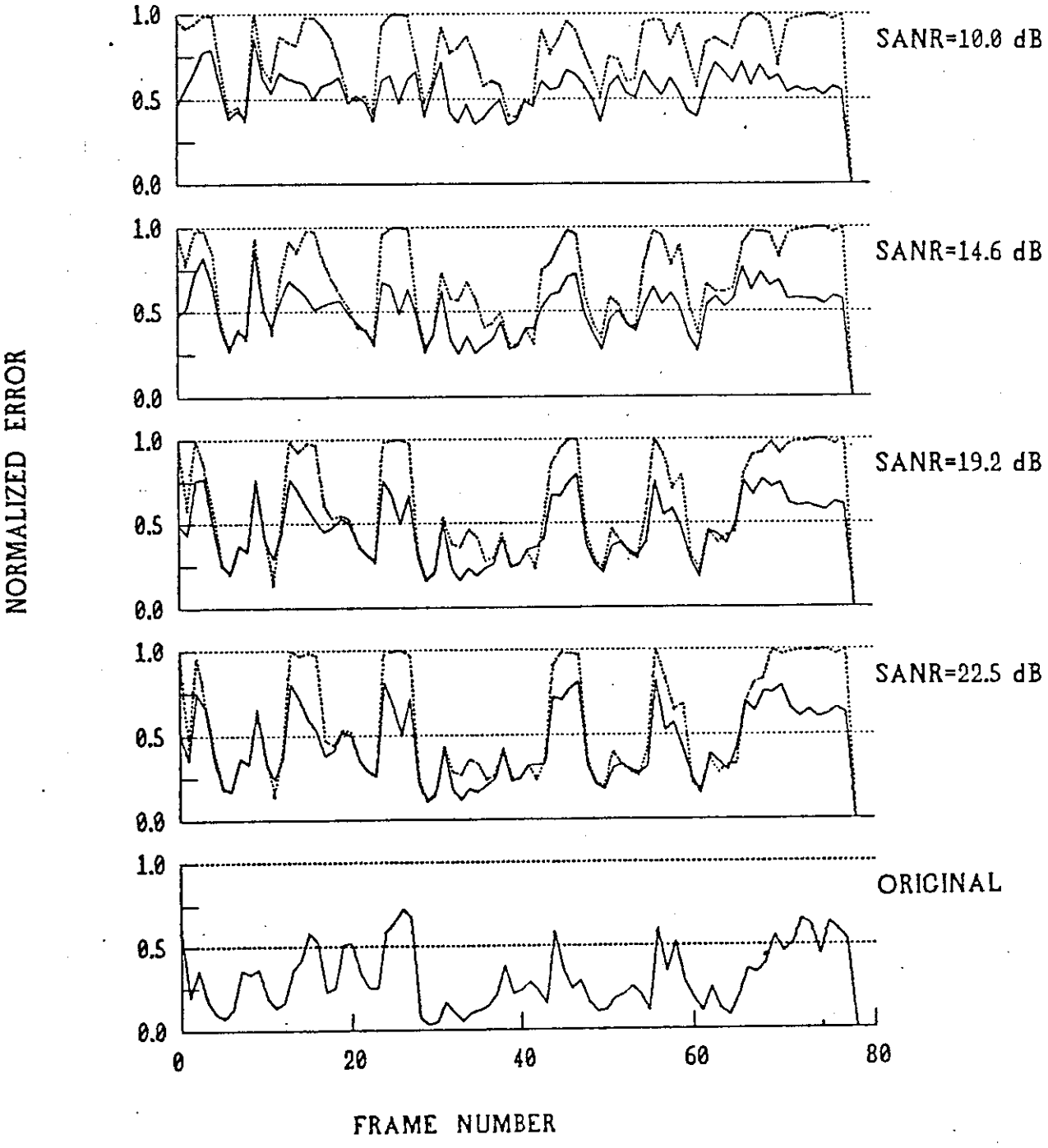

Fig. 25 Normalized LP error (η_m) contour for original data, additive noise distorted data and modified original data (shown by dotted line)