CMU-CS-78-123

PERFORMANCE OF HARPY SPEECH RECOGNITION SYSTEM FOR SPEECH INPUT WITH QUANTIZATION NOISE

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

May 1978

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

ABSTRACT

One of the major problems of a speech processing system is the degradation in performance it suffers due to distortions in the speech input. One such distortion is caused by the quantization noise of waveform encoding schemes which have several attractive features for speech transmission. The objective of this study is to evaluate the performance of the HARPY continuous speech recognition system when the speech input to the system is corrupted by the quantization noise of an ADPCM (Adaptive Differential Pulse Code Modulation) system. The effect of quantization noise on the segmentation and the estimation of LPC (Linear Predictor Coefficients) based paramenters is studied for different bit rates in the range 20-50 kbs of the ADPCM system and the overall word and sentence recognition accuracies are evaluated. The results indicate that even 2-bit ADPCM (corresponding to 20 kbs) speech does not cause significant degradation in performance. The results are explained on the basis of changes produced by the quantization noise in spectral shape and LPC distance.

I. INTRODUCTION

Waveform encoding techniques are generally adopted for efficient transmission of speech information over digital channels. In these cases the signal is corrupted with the quantization noise introduced by the coding scheme. Although many low bit rate schemes have been found to yield perceptually acceptable speech[1], the effect of the accompanying quantization distortion on the performance of speech processing systems such as speech and speaker recognition systems has not been reported. The objective of this paper is to investigate this problem.

The speech processing system considered for investigation is the Harpy continuous speech recognition system developed at Carnegie-Mellon University[2]. We consider the model of Harpy system designed for a 1011 word AI abstract retrieval task. In the system the syntactic, lexical and word juncture knowledge are combined together into one integral network representation. The network consists of a set of states and inter-state pointers. Each state has associated with it phonetic, lexical and duration information. The pointers indicate what states may follow a given state. The initial and final states indicate the beginning and ending points of all utterances respectively. The network is thus a complete (and pre-compiled) representation of all possible pronounciations of all possible utterances in the task language. The recognition process is based on the locus model of search in which all but a narrow beam of paths around the most likely path through the network are rejected.

Recognition process in the Harpy system is as follows: Speech data is sampled at 10kHz and digitized to 9 bits/sample. The sampled data is segmented into acoustically similar sound units based on analyses performed on successive 10 msec segments using Itakura distance metric[3]. A more recent version of the system incorporates segmentation procedure based on ZAPDASH (Zerocrossings And Peaks of Differenced And SmootH waveform) parameters[4] which reduce the computation time for segmentation. Autocorrelation and linear prediction coefficients (LPC) are extracted from the center 10 msec portion of each segment. The segments are then mapped to the network states based on a distance[3] match between the LPC data of the segments and of stored templates. The mapping scheme used is a modified graph search in which heuristics are used to reduce the number of paths that are checked.

As can be seen, any distortion in the input speech can affect at several stages in the recognition process. The segmentation procedures are likely to produce segment boundaries for an utterance different from those in the undistorted speech. The parameters extracted from the segments will also be different and hence a set of templates different from the original ones will be produced. Finally, the distances used for labeling may also be affected causing difficulty in matching the segments to proper network states.

II. GENERATION OF DISTORTED SPEECH DA1A

To study the above mentioned effects on the overall recognition performance of Harpy system we consider quantization noise produced in an ADPCM scheme. The distorted speech is generated as shown in Fig. 1. The scheme uses a feedback adaptive quantization and time invariant first order predictor. Variance adaptive quantization is provided by observing the statistics of the quantizer output and the specification of a corresponding optimum step-size Δ_{opt}. The variance is computed over 64 samples. The following equations define the differential coding[5]:

x_N Input speech samples

E_n Prediction error samples

x_{nq} Quantized input speech samples

E_{nq} Quantized error samples

B Bits per sample

$$E_n = x_n - A_1 \cdot x_{(n-1)q}$$

$$x_{nq} = A_1 \cdot x_{(n-1)q} + E_{nq}$$

$$\Delta_{opt} = K_{opt} \left[\frac{\sum_{n=2}^{N} (x_{nq} - A_1 x_{(n-1)q})^2}{N-1} \right]^{1/2}$$

where $A_1 = 0.875$ and K_{opt} for different values of B are as shown in Table 1.

Table 1. Design values for ADPCM scheme (from Ref. 5).
(sampling frequency = 10 kHz)

Bits per sample B	2	3	4	5
Bit rate I (kBPS)	20	30	40	50
K_{opt}	.996	.586	.335	.225

III. RECOGNITION ON HARPY SYSTEM

Speech data consisting of 55 sentences for training and 20 sentences for testing was recorded using a close speaking microphone. The signal was sampled at 10 kHz sampling rate after passing through a pre-filter (85-4500 Hz). The samples were digitized and stored as 9 bits per sample. ADPCM speech data was generated from these stored samples for the four cases listed in Table 1.

The phone templates for the ADPCM data are generated as follows: The Harpy system is run in a forced recognition mode with a previously generated set of templates for the undistorted speech data. This produces a parsing of the phones to acoustic data. The parsings are used to locate the autocorrelation data for averaging to generate templates for each phone. The averaged templates are tuned further by rejecting the autocorrelation sets that do not fall within ± 1.2 σ (σ is the standard deviation) from the average and computing the average of the remaining sets.

The Harpy system is finally run for recognition of both the training and test data sets. The recognition scores were obtained for both the original and ADPCM data using their respective tuned templates. The overall recognition results are summarized in Tables 2 and 3.

TABLE 2 RECOGNITION RESULTS FOR ORIGINAL AND ADPCM (B=2) DATA

data	word recognition	sentence recognition
ORIGINAL		

Training	98.2(112 114)
	94.0(189 201)
		90.5(19 21)
		88.2(30 34)

Test	92.2(71 77)	90.0(18 20)
ADPCM		

Training	95.6(109 114)
	96.0(193 201)
		81.0(17 21)
		91.2(31 34)

Test	97.4(75 77)	95.0(19 20)

TABLE 3 RECOGNITION RESULTS ON TEST DATA

data	word recognition	sentence recognition
original	92.2(71 77)	90(18 20)
ADPCM B=2	97.4(75 77)	95(19 20)
ADPCM B=3	97.4(75 77)	95(19 20)
ADPCM B=4	94.8(73 77)	90(18 20)

IV. RESULTS AND DISCUSSION

The results in Tables 2 and 3 show that the Harpy system performs equally well for ADPCM speech even with 2 bit coding. This may be due to the fact that the system tunes the templates for each kind of data. Moreover the system uses several sources of knowledge and heuristics to take care of sources of variability such as speaker, noise and distortion. However, if higher accuracy or larger vocabulary systems are built using the finer details of the acoustic data, then the recognition accuracy with distorted speech may not match the performance with the undistorted data.

One of the reasons for obtaining similar recognition performance with distorted and undistorted speech data is probably because most of the spectral information needed for generating phone templates is preserved in the distorted version. Although there is a change in the spectral characteristics of phonemes, as evidenced by the LPC distances, the relative spectral variations among phonemes must have been preserved even in the presence of quantization noise. We have investigated this aspect by observing the short time smoothed spectra of different speech segments and the distance between them. Figs. 2 and 3 show spectra for two different segments of speech for the four types of ADPCM. In each case the smoothed spectrum (dotted line) of the original data segment is also shown for comparison. It is interesting to note that spectral differences caused by the quantization noise are mainly in the low amplitude regions of the spectrum. The significant formant information is mostly retained even for the lowest bit rate (B = 2) ADPCM speech.

LPC distance[3] contours between the original and the ADPCM speech for the utterance "PLEASE HELP ME" is shown in Fig. 4. As expected, the distance between the lowest bit rate ADPCM data and the original is the largest. In order to see how well the relative spectral differences are maintained, the distance contours obtained by comparing adjacent frames is plotted in Fig. 5. It can be seen that in this one frame shift the relative spectral variations are preserved although the absolute distance is smaller for the distorted data.

V. CONCLUSIONS

Speech recognition performance by the Harpy system is not affected significantly by the quantization noise of ADPCM speech. This is probably due to the fact that the system uses several sources of knowledge. Moreover, the system tunes the templates for each kind of data. We have observed that, although the spectral shape is altered due to ADPCM coding, the relative spectral differences among phonemes are preserved as demonstrated in the LPC distance contours. However, if the system is to be designed for higher accuracy or for larger vocabulary, then the finer details of acoustic data may be needed to realize the desired objective. In such a case the performance with distorted speech input may not achieve a comparable recognition performance.

REFERENCES

1. N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM Quantizers," Proc. IEEE, vol. 62, May 1974, pp. 611-632.

2. B. T. Lowerre, The Harpy Speech Recognition System, Ph.D. Dissertation, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1976.

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

4. H. Goldberg, D. R. Reddy and G. Gill, "The ZAPDASH parameters, feature extraction, segmentation, and labelling for speech understanding systems," in CM 77 Su.

5. M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantization noise or additive white noise," IEEE Trans. Acoust. Speech and Signal Processing, vol. ASSP-24, December 1976, pp. 488-494.

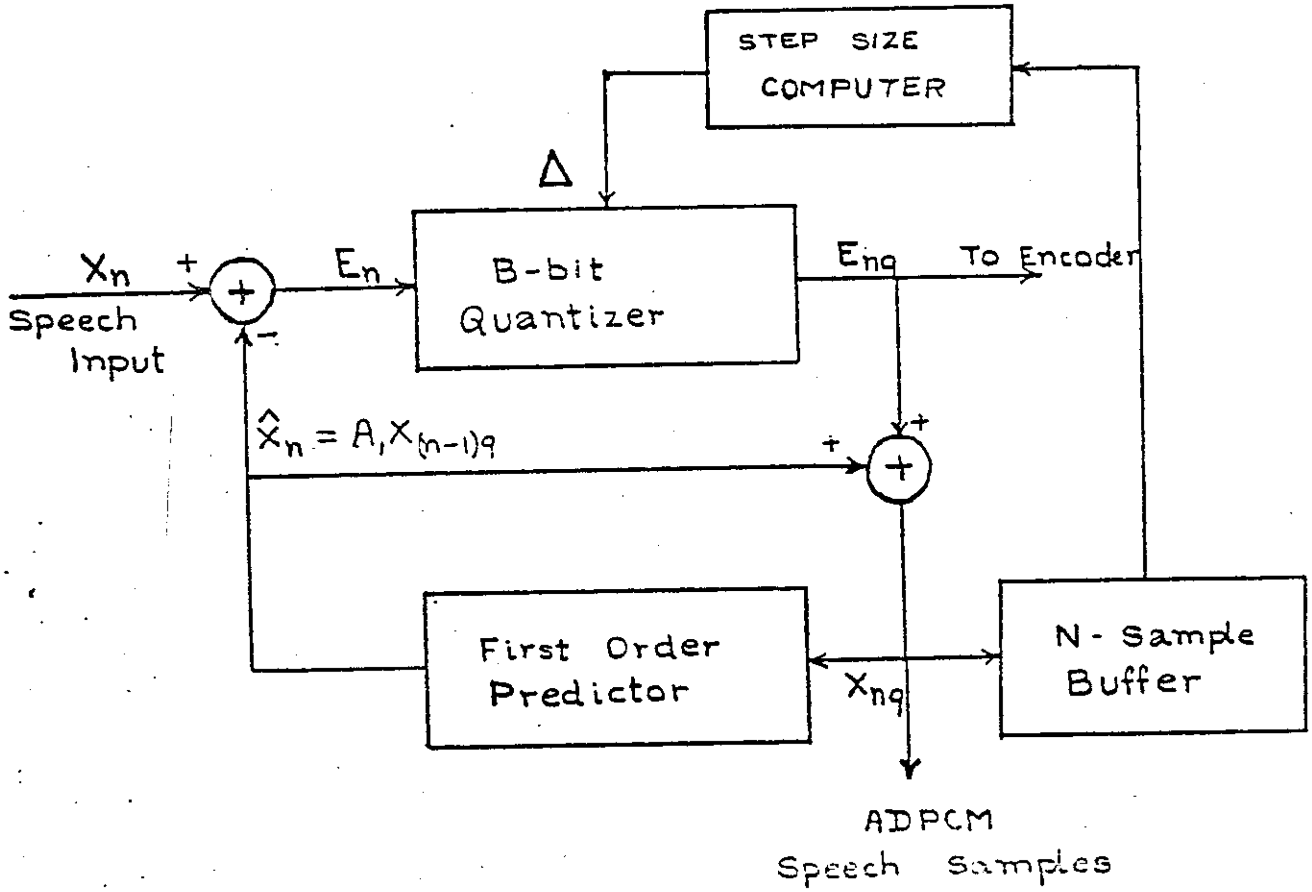

Fig.1. Generation of ADPCM Data

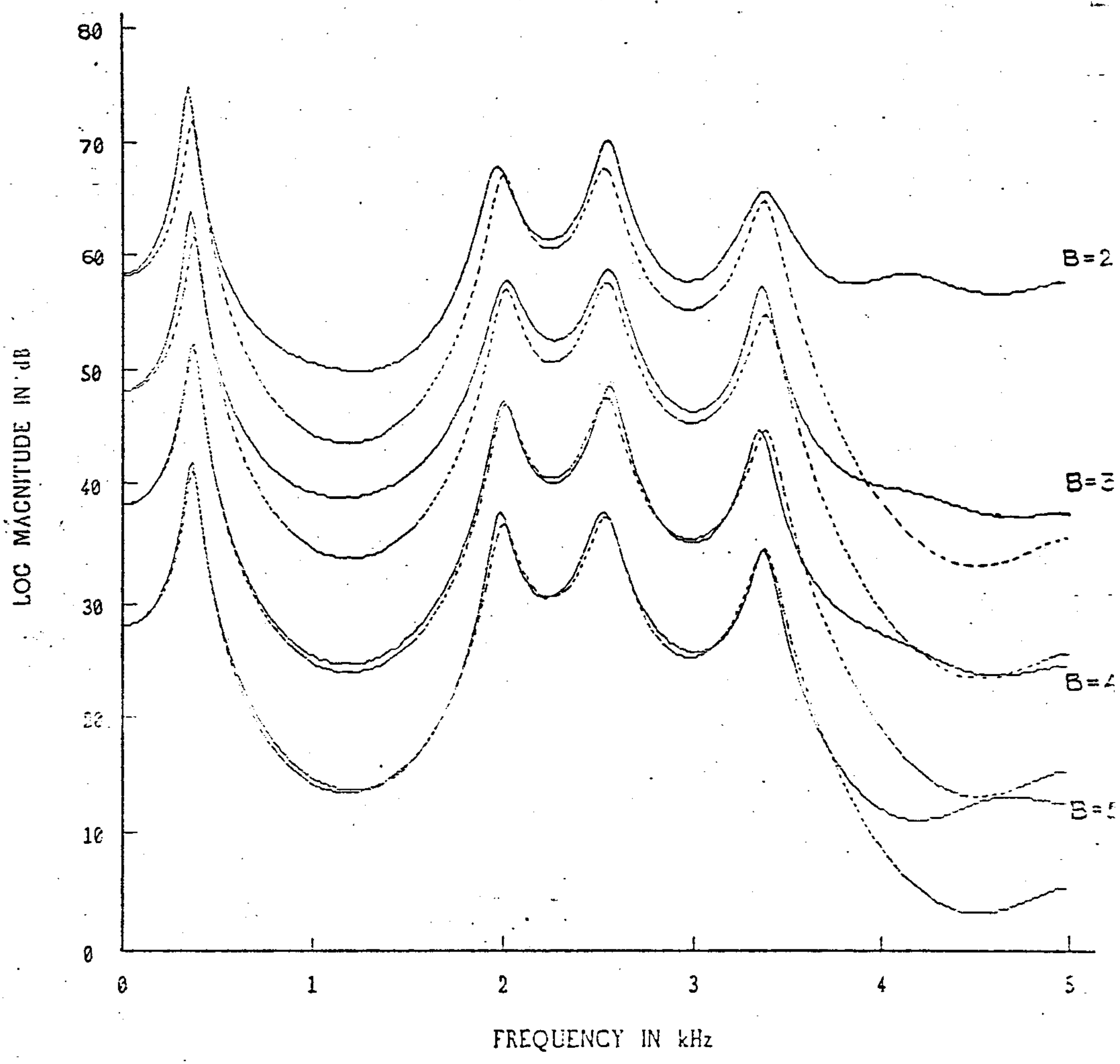

Fig. 2. LP Smoothed Spectra for a vowel segment of ADPCM data

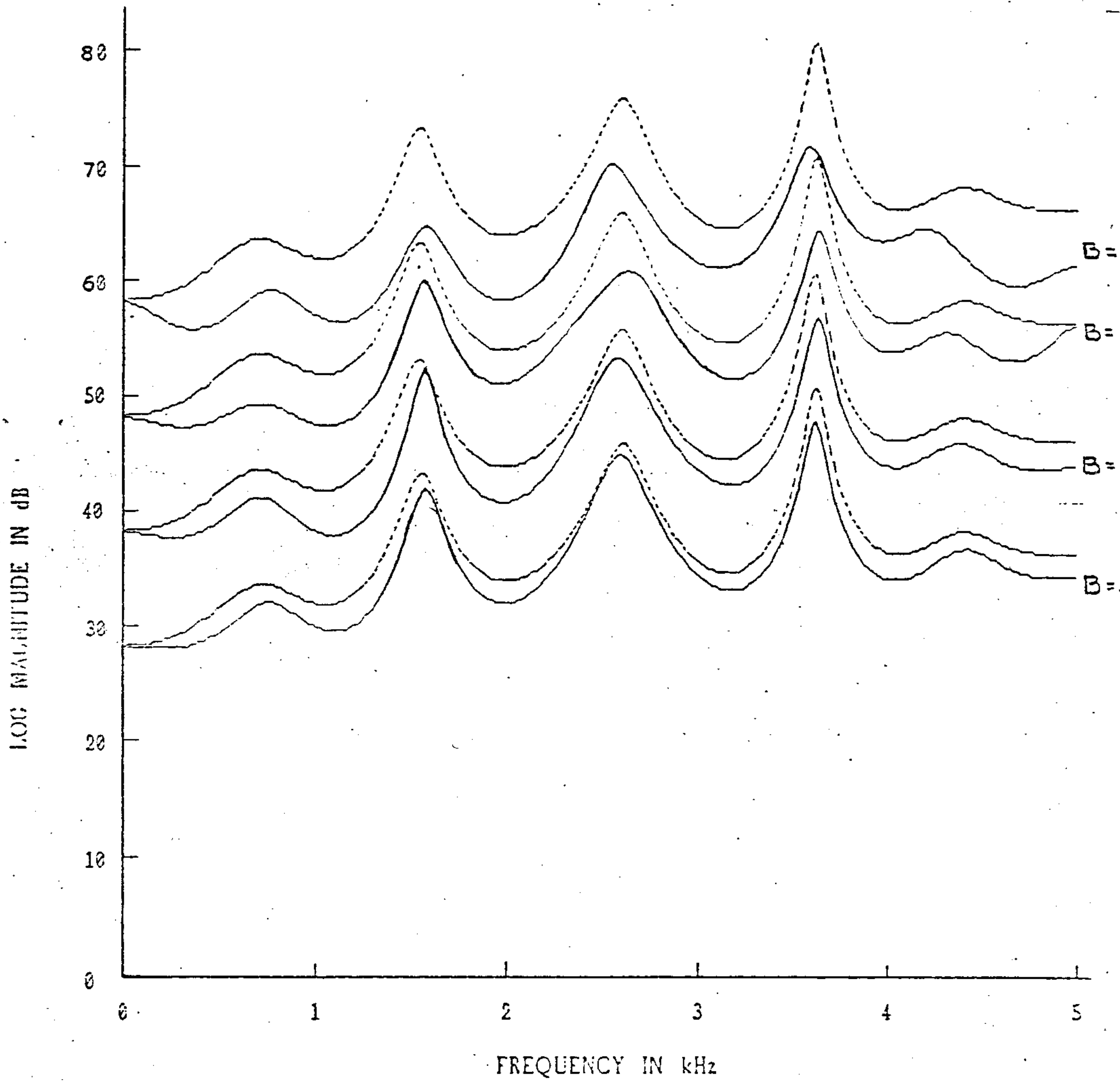

Fig.3. LP Smoothed Spectra for an unvoiced segment of ADPCM data

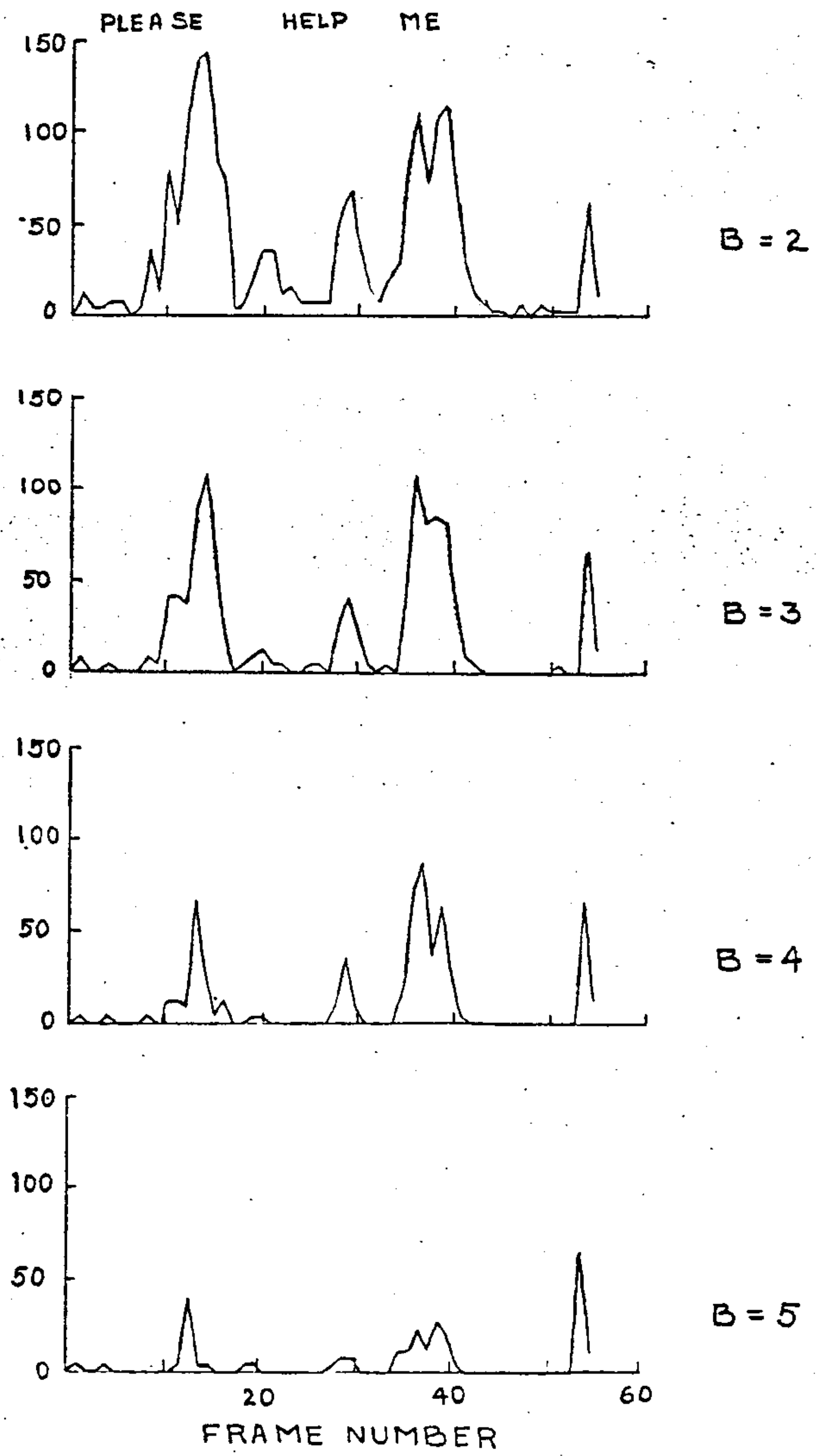

Fig.4. LPC distance contours between original and ADPCM data

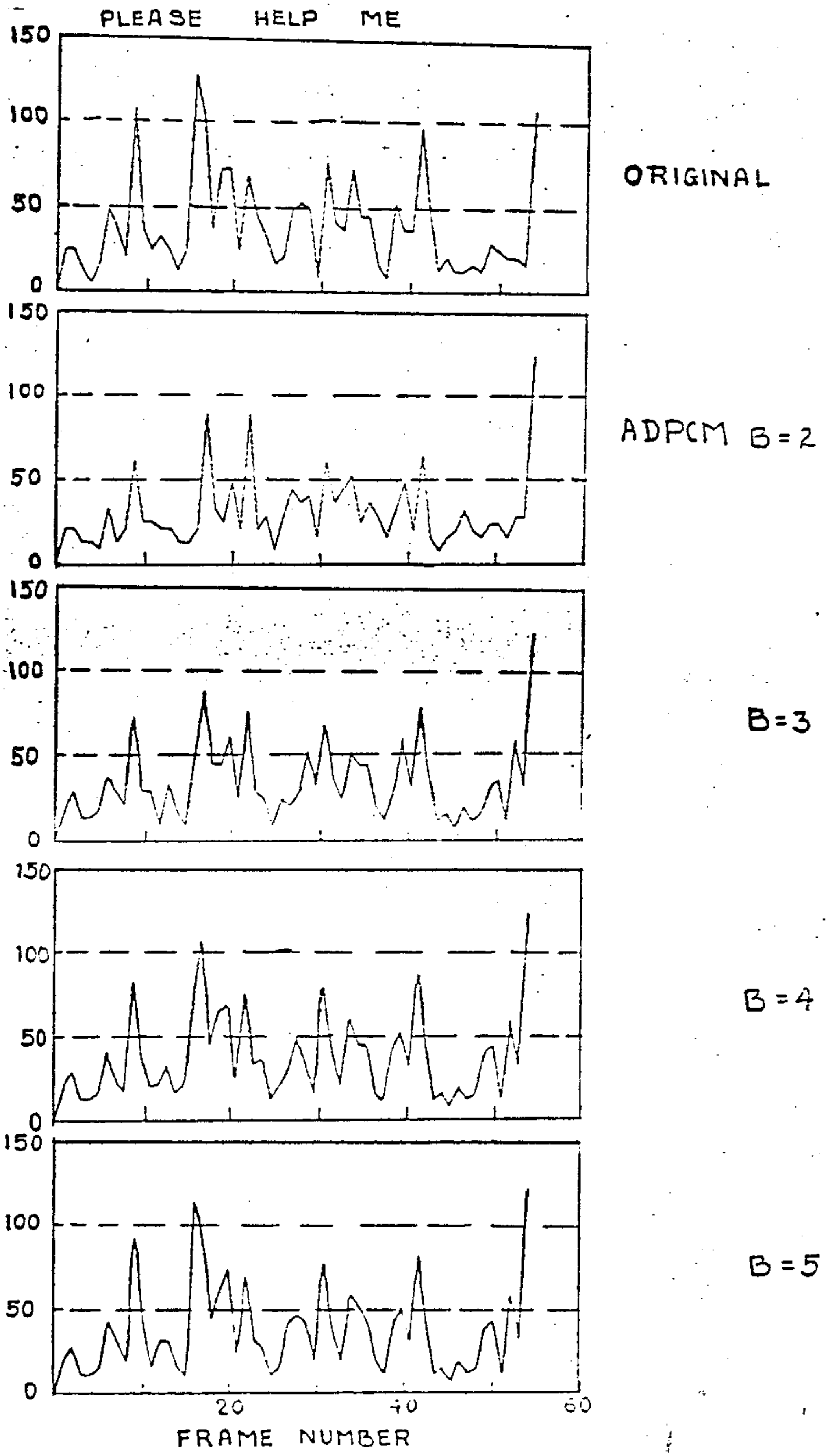

Fig.5 LPC distance contours between adjacent frames