THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM

By B. McDERMOTT, C. SCAGLIOLA, and D. GOODMAN

(Manuscript received November 9, 1977)

An experiment has been performed to study the perceptual characteristics of speech processed by adaptive differential PCM. We created 18 three-bit and four-bit coders spanning a wide range of quantizer adaptation parameters. Subjects judged differences between coders and rated the quality of each coder individually. The difference data reveal three important perceptual characteristics: overall clarity, signal vs. background degradation, and rough vs. smooth impairment. These characteristics are strongly correlated with coder design parameters and objective performance measures. Overall subjective quality is well predicted by segmental signal-to-noise ratio and even better by a linear combination of measures of granular distortion and overload distortion.

I. INTRODUCTION

Speech signal processing systems are susceptible to a variety of audible impairments often classified with words like "distortion," "noise," "echo," and "sidetone." These categories are themselves subdivided: for example, "linear" and "nonlinear" distortion, "white" noise, "impulsive" noise, "speech-dependent" noise, etc. When the type of system is familiar to a large body of listeners, the application of these names becomes standardized and a language exists for describing the quality of specific implementations. With new systems, however, the types of degradation are often not known *a priori*, and special effort is required to identify them and to relate them to physical characteristics of the system.

For example, experiments on PCM (pulse code modulation) have identified peak clipping, granular quantizing noise, and bandlimiting as important audible degradations.[1,2] In PCM, there are relatively few design parameters, and each of these impairments can be related to one of them: peak clipping to quantizer overload point, granular noise to step size, and bandlimiting to sampling rate.

In ADPCM (adaptive differential PCM), a coding method that appears promising for a number of practical applications, the situation is more complicated. Here each design parameter has interrelated effects on several types of degradation, and the perceptual correlates of a particular design are hard to predict. In ADPCM, the step size and overload point vary with time, producing a dynamic mixture of overload and granularity that depends on the adaptation mechanism. One can identify in the quantized waveform two types of overload: overload that causes clipping of stationary inputs and in addition, overload due to slow quantizer response to increases in short term (syllabic) signal level. Moreover, a mathematical study[3] has identified two separate aspects of adaptive quantizer performance: static (response to constant-level inputs) and dynamic (response to changes in input level). However, it is by no means evident or even likely *a priori* that these mathematically separable characteristics are perceived separately.

To investigate perceptual characteristics of speech processed by ADPCM we conducted an experiment that is summarized in the next section. The following five sections provide details of the coding method, objective performance measures, the experimental design, and analyses of subjective and objective measurement data. Section VIII discusses the implications of the principal findings.

II. SUMMARY

High-quality digital recordings of speech samples from four talkers (two male and two female) were processed according to 18 different ADPCM coding schemes on a digital computer. The coders incorporate all combinations of two bit rates, three load constants, and three time constants. The data obtained from the experiment consisted of two types: objective measurements and subjective judgments of the processed speech. With these two types of data we addressed the following questions:

(*i*) What are the perceived characteristics of speech processed by ADPCM?

(*ii*) How are these characteristics related to subjective judgments of quality?

(*iii*) What is the relationship between objective performance measures and the perceptual features?

(*iv*) What is the relationship between objective performance measures and judgments of circuit quality?

(*v*) What combinations of design parameters produce coders within a given quality range?

The analyses of the data indicate the following answers to each of the above questions:

(*i*) Listeners perceive three distinct characteristics of the processed speech: (*a*) the overall clarity, (*b*) the kind of degradation that reduces the clarity, namely, whether the degradation is signal distortion or noise or both, and (*c*) the nature of the signal distortion and/or the nature of the noise.

(*ii*) Quality judgments are correlated with all of these subjective variables. The overall clarity is by far the strongest correlate.

(*iii*) Signal-to-noise ratio, measured segmentally, SNR_{seg}, is a good predictor of the overall clarity. The log of the ratio of attack to recovery speed, $\log A/R$, is a good predictor of the mixture of signal distortion and background noise. The log of the attack time, $\log T_a$, predicts the kind of signal distortion and/or the kind of noise.

(*iv*) SNR_{seg} is a very good predictor of quality judgments, while SNR measured in the traditional manner is a very poor predictor of quality. A linear combination of probability of overload, P, and segmental signal-to-granular-noise ratio, SNRG_{seg} is an even better predictor of quality than SNR_{seg}.

(*v*) By applying the prediction equations to coders with design values intermediate to those of the experiment, we show the combinations of load constant and time constant at each bit rate that would be judged about equal in quality. Those that would be rated almost as highly as the best coder cover a wide range of design parameters.

III. CODER DEFINITIONS

Figure 1 is a block diagram of an ADPCM coder-decoder. In the absence of transmission errors, the sequence of received samples $r'(k)$, is identical to the quantized approximation sequence $r(k)$. In our experiment $s(k)$ was a digital speech signal represented in a 12 bit, 8 kHz format and the coders were realized in software on a Data General Eclipse computer.

The conversion from 12-bit PCM to 3-bit or 4-bit ADPCM is performed according to the algorithm described by Castellino et al.[4] In all of the coders the predictor is a two tap transversal filter with coefficients 1 and -0.5 so that the relationship of approximation signal, $r(k)$, to quantizer output, $d(k)$, is

$$r(k) = d(k) + r(k-1) - 0.5r(k-2). \tag{1}$$

Signal-level estimation. The step size, $\Delta(k)$, which is derived from the sequence of quantized prediction error samples $d(k)$ or equivalently from the transmitted code words, $I(k)$, is proportional to an estimate of the mean absolute value of the quantizer input, $e(k)$. The estimate at time k, $\sigma(k)$, is an exponentially weighted sum of quantizer output magnitudes. It is computed recursively as

$$\sigma(k) = \alpha\sigma(k-1) + (1-\alpha)|d(k-1)|. \tag{2}$$

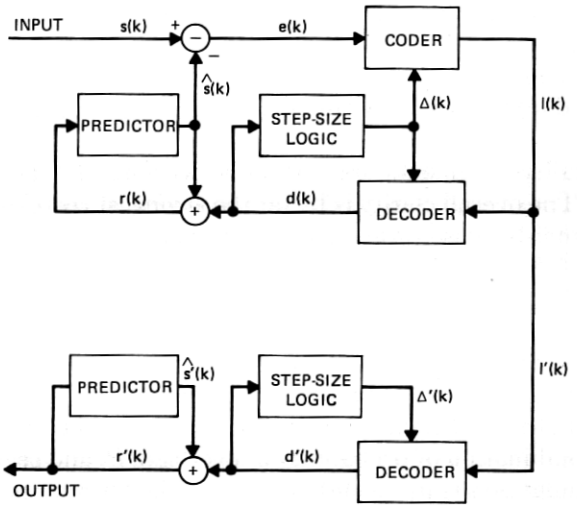

Fig. 1—Block diagram of ADPCM coder-decoder.

Here, the parameter α ($0 < \alpha < 1$) determines the speed of response of the quantizer to changes in input level. A low value of α provides a fast response; $\alpha \approx 1$ is associated with a slow response. A compromise between static and dynamic performance objectives is required in selecting α. A high value ($\alpha \approx 1$) provides more accurate "steady-state" tracking of a constant signal level than does a low value.

If the signal level suddenly increases, the estimate σ increases with an initial slope proportional to $1 - \alpha$ volts/sample (if σ is measured in volts). In this paper we shall refer to an adaptation time constant, τ sec, that is the reciprocal of this initial slope in response to a unit step. It is defined by

$$\tau = \frac{T}{1 - \alpha} \text{ sec}$$

where T is the sample period. For 8 kHz sampling

$$\tau = \frac{0.125}{1 - \alpha} \text{ msec.}$$

To incorporate a perceptibly wide range of signal conditions, we have selected, after informally listening to a large number of coders, 3 values of α for the experiment. They are $\alpha = 1/2, 31/32, 255/256$ with corresponding time constants: $\tau = 0.25, 4, 32$ msec.

Quantizer loading. The quantization step-size $\Delta(k)$ is proportional to the signal-level estimate, $\sigma(k)$:

$$\Delta(k) = C\sigma(k) \tag{3}$$

where the load constant, C, determines in steady-state (fixed signal level) the mixture of granular noise and overload distortion in $d(k)$. A relatively high value of C produces a large average step size and causes granularity to be the principal distortion component. With a very low value of C, overload predominates.

For a given number B bits/sample, we define a nominal load constant, C_0. C_0 is the step size which produces minimum mean square error in a fixed quantizer processing a signal that has Laplacian probability density with unity average magnitude. In fact, it has been noted that the shape of the PDF of the compressed signal $e(k)/\sigma(k)$ is between a Gaussian and a Laplacian function, and more near to the Laplacian one for higher time constants.[5] For $B = 2, 3, 4, 5$ bits, $C_0 = 1.53, 1.03, 0.65, 0.40$, respectively. We define a relative load constant for each quantizer to be

$$L = 20 \log_{10} \frac{C}{C_0} \text{ dB}.$$

After listening informally to speech processed by a variety of coders, we chose for the experiment three relative load constants, $L = -10, -4, 4$ dB.

It has been shown[6] that this adaptive quantizer is a special case of the one with multiplicative step size changes[7]:

$$\Delta(k + 1) = M[I(k)]\Delta(k).$$

The multipliers associated with code words $I(k) = \pm 1, \pm 2, \ldots \pm 2^{B-1}$ are

$$M[I(k)] = \alpha + (1 - \alpha)C(|I(k)| - 0.5).$$

Dynamic and static behavior. The dynamic behavior of the coder can be described by two characteristics: the attack and recovery speeds. The attack speed is defined as the step size increase (in dB) per unit time when the signal level suddenly changes from a very low value to a very high value. The recovery speed is defined as the step size decrease (in dB) per unit time when the signal level suddenly falls. The attack and recovery speeds can be computed from the largest and smallest multipliers[3]:

$$\nu_a = \frac{20}{T} \log M(I_n) \frac{\text{dB}}{\text{sec}}$$

$$\nu_r = \frac{-20}{T} \log M(I_1) \frac{\text{dB}}{\text{sec}}$$

where T is the sampling time, and $n = 2^{B-1}$. A small attack speed will produce slope overload distortion, while a small recovery speed will result

in greater granular distortion. The static behavior is also affected by attack and recovery speed: a high attack and low recovery speed result in a step size that is higher on the average than that resulting from a slow attack and fast recovery. A very good indicator of static performance is the attack to recovery ratio:

$$A/R = \frac{\nu_a}{\nu_r} = \frac{\log M(I_n)}{-\log M(I_1)}.$$

Attack time, which we have found to be strongly correlated with the type of distortion or the type of noise produced by a coder, is the reciprocal of attack speed:

$$T_a = \frac{1}{\nu_a}.$$

Summary of conditions. The experiment includes coders with 3 variable design parameters: B bits/sample, τ msec response time, and L dB relative load constant. The 18 coders comprise all combinations of $B = 3, 4$; $\tau = 0.25, 4, 32$; $L = -10, -4, 4$.

IV. OBJECTIVE MEASURES

Our aims include exploration of the relationships between perceived characteristics of the processed speech and objectively measurable quantities. To investigate these relationships, we have computed several objective performance indices for each processed utterance. The measures are defined as follows:

Total signal-to-noise ratio.

$$\text{SNR} = 10 \log \frac{\Sigma s^2(k)}{\Sigma[s(k) - r(k)]^2}.$$

Here k ranges over all samples in the utterance, and $r(k)$ is defined as the best estimate of $s(k)$.

Granular signal-to-noise ratio.

$$\text{SNRG} = 10 \log \frac{\Sigma s_g^2(k)}{\Sigma[s_g(k) - r_g(k)]^2}$$

where k ranges over all samples in the utterance and the signals $s_g(k)$ and $r_g(k)$ are defined only when the quantizer is not overloaded; that is, when the quantization error is less than one-half the step size:

$$s_g(k) = s(k); r_g(k) = r(k) \quad \text{if} \quad |s(k) - r(k)| \leqslant \frac{\Delta(k)}{2}$$

$$s_g(k) = r_g(k) = 0 \quad \text{if} \quad |s(k) - r(k)| > \frac{\Delta(k)}{2}.$$

Percent of samples overloaded.

$$P = 100\left[1 - \frac{\sum_{k=1}^{N} s_g(k)/s(k)}{N}\right]$$

where N is the total number of samples in the utterance.

Total segmental signal-to-noise ratio. This is a measure proposed by Noll[8] as a more relevant index of speech quality than SNR:

$$\text{SNR}_{seg} = \frac{1}{M}\sum_{m=0}^{M-1} 10 \log \frac{\sum_{j=1}^{128} s^2(j + 128m)}{\sum_{j=1}^{128} [s(j + 128m) - r(j + 128m)]^2}$$

Here the utterance is divided into segments each containing 128 samples (16 msec) and the signal-to-noise ratio in each segment is measured in dB. The average of these measures over the M segments in the utterance is SNR_{seg}.

Granular segmental signal-to-noise ratio.

$$\text{SNRG}_{seg} = \frac{1}{M}\sum_{m=0}^{M-1} 10 \log \frac{\sum_{j=1}^{128} s_g^2(j + 128m)}{\sum_{j=1}^{128} [s_g(j + 128m) - r_g(j + 128m)]^2}$$

Here the same procedure as for SNR_{seg} is applied to samples $s_g(k)$ and $r_g(k)$.

V. TESTING PROCEDURE

Digital recordings* of 10 sentences spoken by each of 4 talkers (2 male and 2 female) were processed by each of the 18 coders. The processed sentences were equalized to the same mean power to eliminate level differences due to quantizer overloading and thereby minimize differences in subjective loudness. Four analog test tapes, each containing the 153 possible pairs of coders, were prepared from these recordings. The speech samples in a pair of conditions were the same sentence by the same talker. The talkers and sentences were assigned to coder pairs so that they occurred as equally as possible on a tape. A pair of coders processed a different talker and sentence on each tape and the order of

* The source speech was the set of digital tape recordings used in a previous experiment on PCM.[1]

presentation within the pair was reversed on half of the tapes. The coder pairs appeared in a different (random) order on each tape.

Students from the junior and senior classes of local high schools served as paid subjects. They listened to the processed speech over Pioneer SE700 earphones at 80 dBSPL while seated in a double-walled sound booth with frequency-weighted room noise introduced at a level of 50 dBA.

Dissimilarity judgments. In the first experiment, 17 subjects (3,4,5,5 per random order) judged the pairs of conditions. They were told to use the numbers from 0 to 9 to indicate how different the speech sounded over each pair of coders, using a 0 for no difference, a 9 for very different, and the numbers between 0 and 9 for intermediate differences. Before the test session began they judged 6 pairs, for practice, that represented the expected range of differences.

Preference judgments. In the second experiment, 16 different subjects (4,5,4,3 per random order), also junior and senior high-school students, listened to the same tapes containing the pairs of coders. However, this time the subjects were instructed to indicate which condition of each pair they would find more acceptable for listening to speech.

Rating judgments. In the third experiment, subjects judged the quality of the coders individually. Eight audio tapes were prepared, each containing 36 sentences. Tapes 1–4 had the processed speech (played through the 18 coders) of one male and one female talker. The other two talkers appeared on tapes 5–8. The stimuli on tapes 1–4 appeared in different (randomized) orders. The same 4 orderings were used for tapes 5–8. The sentences occurred as equally as possible on each tape.

In this experiment the subjects were asked to rate the quality of the 36 conditions according to the adjectives: excellent, good, fair, poor, unsatisfactory. Their answer sheets contained 36 rows of short lines separated into 9 columns. The odd columns were labeled with the adjectives and the even ones unlabeled, allowing the subjects to check intermediate ratings if they chose to do so. On half the answer sheets the order of the labels were reversed. Tapes 1–4 were presented to the 17 listeners of the first experiment in a short session that took place 5 minutes after the completion of the difference judgments. Tapes 5–8 were presented to the 16 listeners of the second experiment 5 minutes after the completion of the preference judgments.

VI. INITIAL DATA REDUCTION

The experiment was performed to provide information about relationships between ADPCM coders and it is expected that for the most part differences in listener responses are due to coder differences. The experiment was designed to cause other sources of variability to be mutually cancelling in the average data for each coder. Sources of extraneous

variability are: differences in the way listeners use the response scales, differences in speech material, and effects of presentation order. Before aggregating the data for individual coders it was necessary to assess the importance of each of them.

Difference judgments. The variability due to listener differences is revealed by the correlation coefficients of pairs of subjects who heard the same tape. These coefficients have a mean value of 0.61 and standard deviation of 0.09, indicating substantial agreement. The effects of presentation order and talker were tested by means of an analysis of variance which showed that the responses to the 4 random orders were not significantly different. The variability due to the different talkers was significant at the 0.05 level, but accounted for only 1 percent of the total variance. This variability is due to the fact that speech of female talkers was rated differently from speech of male talkers. There was no significant difference in the ratings of talkers of the same sex.

The important variability in the difference data can therefore be attributed to coder differences, and to assess these differences, we normalized the 153 responses of each subject to zero mean and unity standard deviation. The averages, across the 17 subjects, of the normalized responses were the elements of a dissimiliarity matrix which was analyzed according to the MDSCAL[9–13] procedure.

MDSCAL locates points, representing the stimuli, in a multidimensional space so that the distances between the points are monotonically related to the judged differences. Because the dimensionality of a solution is specified as input, successive solutions of increasing dimensionality are usually computed. Then, the stress values (essentially the root mean square error) and the interpretability of each solution are used as criteria for deciding upon the smallest number of dimensions that are needed to explain the data. The stress values give a measure of how well the distances in the solution spaces correspond to the reported differences among the coders. Solutions in 1, 2, and 3 dimensions for the difference judgments among the 18 coders had stress values of 0.25, 0.11, and 0.07, respectively. The large decrease in stress between the 1 and 2 dimensional solutions indicates that at least 2 dimensions are needed to account for the data. Although a 3-dimensional solution accounted for only a small additional decrease in stress, it offered an enhanced interpretation of the subjective space. (See Section VII.)

Preference judgments. The preference data were analyzed according to MDPREF,[14,15] a factor analytic procedure that measures the variability in preference among the subjects. The proportion of the total variance contributed by each factor is related to the agreement among the subjects on the relative importance of different characteristics of the stimuli. In the solution for the preference judgments of the 18 coders, the first factor accounted for 0.89 of the variance and the second accounted for only an

additional 0.02, indicating strong agreement among the listeners and a single factor solution. Therefore, the values of the points from the one factor solution were used for the scale of preference.

Rating judgments. We computed the correlations of the ratings of each subject with those of each of the other subjects who listened to the same tape. The distribution of the correlation coefficients had a mean of 0.78 and standard deviation of 0.08, again showing a high degree of agreement. Therefore, the mean across subjects of the individual responses, normalized so that the ratings of each subject had zero mean and unit variance, were used in an analysis of variance due to coder parameters, talkers, and random orders. The 3 design variables, load constant, time constant, and bits, were all significant at the 0.05 level. The variability due to the random orders was not significant, but the variability due to the different talkers was significant. As in the difference experiment, the significant talker variability was due to differences between the male and female talkers, accounting for only 1 percent of the variance.

Rating vs. preference. The two types of quality judgments, preference and rating, were obtained so that the two testing methods could be compared. The paired comparison tapes were designed to balance many of the sources of variability that are artifacts of the testing procedure. Each coder was heard an equal number of times with each talker and approximately an equal number of times with each sentence. The order of presentation was reversed on half of the trials and, of course, the relative merit of each coder was ultimately determined by comparing it with every other coder. In the rating judgments, the merit of a coder was determined by one presentation per talker. Ratings assume that the quality represented by the five adjectival categories are not only well defined for each individual, but are essentially the same for all individuals. Although the ratings were normalized before computing the analysis of variance, the more customary procedure is to simply average the original judgments across subjects. Therefore, to compare the results of the rating study with those of the more critical paired-comparison study, the mean across subjects of the original unnormalized ratings were used.

Figure 2 shows a scatter plot of these ratings vs. transformed quality measures from the one factor MDPREF solution of the paired-comparison judgments. (The linear transformation scales the maximum and minimum measures to one and nine, respectively.) As this plot shows, the agreement in ratings for the two methods is extremely high: the correlation is 0.99. Thus it appears that the uncontrolled sources of variability that could contaminate simple rating judgments did not have a strong influence on the variability of these data. The additional experimental effort involved in collecting paired-comparison judgments in order to control this variability did not increase the accuracy.

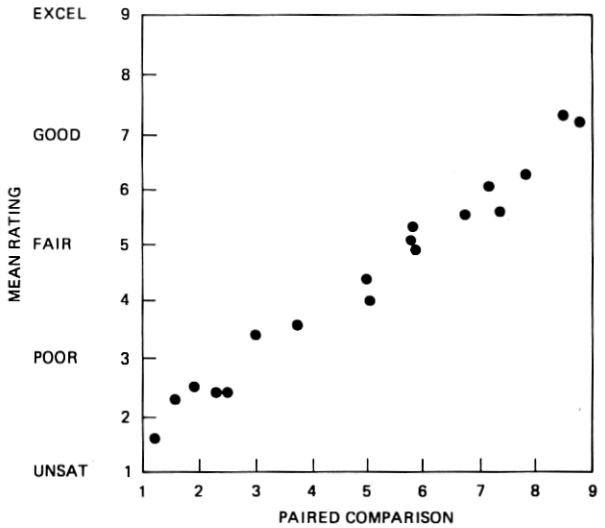

Fig. 2—Relationship between overall evaluation of coders by category ratings and paired comparison preference methods.

VII. RESULTS

The output of MDSCAL is a set of points, representing the coders, in Euclidean space. The inter-point distances are related to the judged differences between coders and the orientation of the points in this space is supposed to reveal the underlying perceptual characteristics of the coders. However, the distances are invariant under orthogonal rotation and it is often the case that a rotation of the MDSCAL coordinates is necessary to interpret the configuration in terms of known coder properties.

In evaluating MDSCAL analyses of varying dimensionality, we concluded that a 3-dimensional geometry would be most informative. We approached the rotation problem by using multiple linear regression procedures to locate vectors in the 3-dimensional space on which the projections of the points are maximally correlated with various objective measures and design parameters. We also located the vector on which the projections of the points are maximally correlated with the average subjective ratings obtained in the third experiment. The vector for each measure of the 18 coders was located independently. Table I displays some of the measures for which vectors were derived and Table II gives the correlations between measurements and the corresponding projections on vectors in the MDSCAL space. These vectors are an aid to the interpretation of the subjective space because they make it possible to relate directions in space to changes in design parameters and performance measures. As a visual aid, the coordinate axes were rotated so that they nearly or exactly coincide with meaningful directions.

Table I

Coder	B	τ ms	L, db	T_a, ms	T_r, ms	A/R	SNR, dB	SNR_{seg}, dB	SNRG, dB	$SNRG_{seg}$, dB	P, %	R
1	3	0.25	4	0.01	0.15	12.63	10.5	8.8	10.5	8.8	0.07	3.4
2	3	0.25	−4	0.03	0.04	1.20	14.8	15.7	20.3	19.0	6.69	5.7
3	3	0.25	−10	0.21	0.03	0.13	2.9	4.5	24.9	23.7	58.28	2.4
4	3	4.00	4	0.11	2.51	23.99	11.1	9.0	11.1	9.2	0.11	3.2
5	3	4.00	−4	0.37	0.68	1.84	13.1	14.6	20.9	19.6	6.15	5.6
6	3	4.00	−10	3.27	0.54	0.17	2.1	4.2	24.3	23.0	57.32	2.5
7	3	32.00	4	0.79	20.13	25.54	11.1	5.5	11.2	5.9	0.60	2.3
8	3	32.00	−4	2.89	5.45	1.89	7.9	11.2	20.7	17.8	13.03	4.4
9	3	32.00	−10	26.16	4.39	0.17	0.7	3.6	22.4	22.9	63.79	1.6
10	4	0.25	4	0.01	0.05	5.32	15.9	14.8	16.0	14.8	0.07	4.9
11	4	0.25	−4	0.02	0.03	1.41	19.2	20.4	24.5	23.1	2.75	7.4
12	4	0.25	−10	0.06	0.02	0.40	8.9	11.7	29.9	28.1	25.74	5.2
13	4	4.00	4	0.08	0.94	12.53	17.1	15.5	17.2	15.7	0.11	5.2
14	4	4.00	−4	0.23	0.57	2.50	17.0	19.7	25.5	23.9	2.59	7.2
15	4	4.00	−10	0.85	0.51	0.59	7.4	10.3	29.8	27.9	24.91	6.2
16	4	32.00	4	0.55	7.60	13.73	17.1	13.0	17.7	13.5	0.44	4.0
17	4	32.00	−4	1.78	4.63	2.60	11.8	16.2	25.5	22.3	6.81	6.1
18	4	32.00	−10	6.78	4.10	0.60	3.3	9.4	28.9	27.1	34.90	3.5

Table II

Objective measure	Corr. with vector values
SNR	0.95
SNR_{seg}	0.96
log A/R	0.96
SNRG	0.94
$SNRG_{seg}$	0.95
P overload	0.98
log T_a	0.92
log T_r	0.85
Rating	0.99

As a further step in interpreting the space, we listened to one of the tapes used in the rating experiment. After hearing each sentence, the three of us independently wrote adjectives to describe the processed speech. Examples of coder descriptions are: "clear, some noise," "slightly muffled, medium noise," "crackling noise," "very hoarse."

After considering several other rotations, we chose the solution displayed in Figs. 3 and 4 as most interpretable because the coordinate axes nearly or exactly coincide with vectors of measurable quantities and the coder descriptions cluster in a meaningful way. With this rotation, the proportions of the total variance accounted for by dimensions I, II, and III are 0.62, 0.19, and 0.19, respectively.

Subjective variables. When the descriptive adjectives were related to the configuration of points on the plane of the first two dimensions, shown in Fig. 3, an interpretation emerged that was reminiscent of a similar analysis in a study of analog circuits.[16] The interpretation of the space in that study indicated that listeners distinguish among the pro-

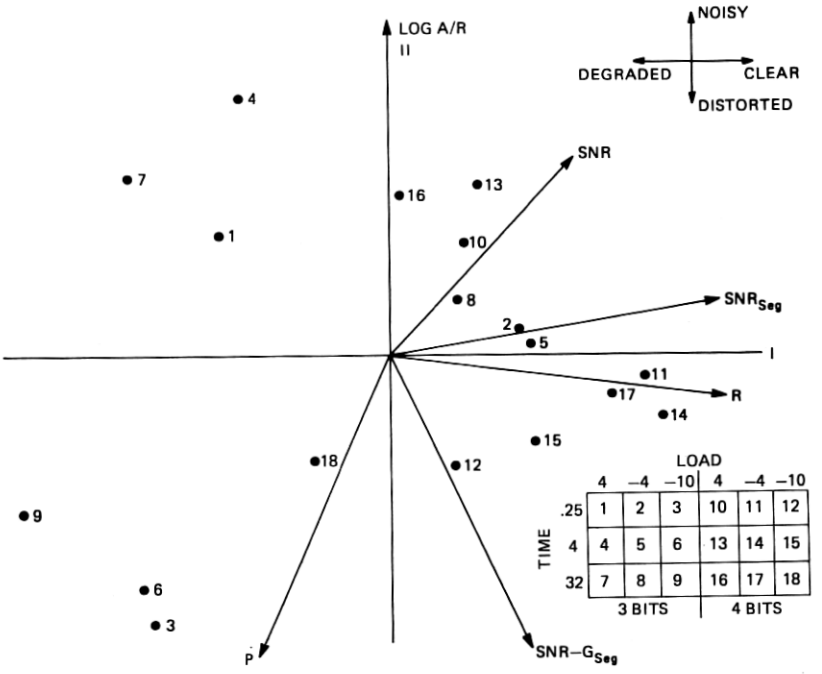

Fig. 3—Projections of points, representing coders, and vectors, representing measures, on the plane of dimensions I and II.

cessed speech samples according to whether the speech is clear or degraded. When degradation is present, they further distinguish between noise in addition to the signal and distortion of the signal itself. The same interpretation applies to the first 2 dimensions in the ADPCM coder space. The coders we described as having clear speech and little noise are high on the first dimension. The coders that were described as noisy, muffled, and hoarse are low on the first dimension, and intermediate amounts of each type of reduction in overall clarity are distributed between these two extremes. Thus, the first dimension appears to represent the overall clarity of the speech.

The plane of dimensions II and III, shown in Fig. 4, identifies the characteristics of the speech that reduce the clarity. The coders that we described as noisy are high on the second dimension and those that we described as muffled or hoarse are low on the second dimension. Thus, the second dimension represents the two kinds of degradations that reduce the overall clarity: background noise and distortion of the speech signal itself. The conditions we described as very muffled sounded as though the speaker had his hand, or some other material object, in front of his mouth, and these conditions are high on dimension III. Those conditions we described as hoarse sounded as though the speaker had

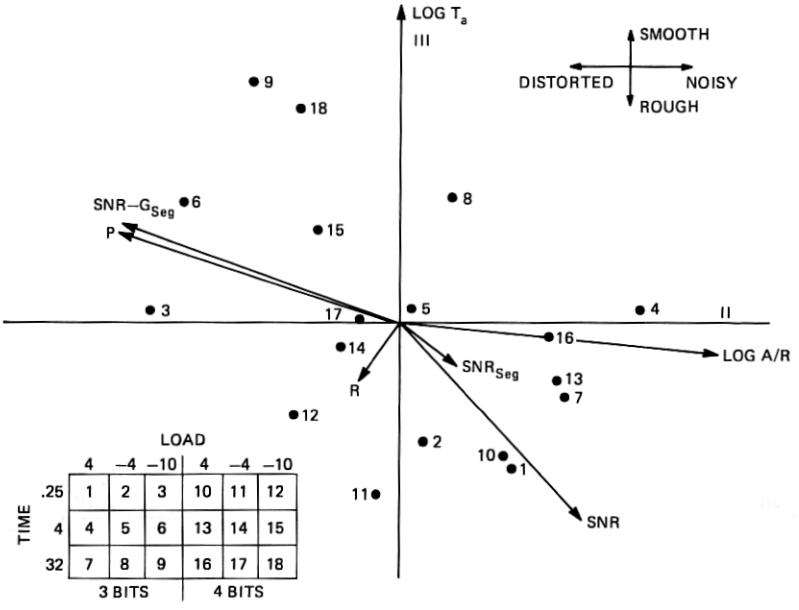

Fig. 4—Projections of points, representing coders, and vectors, representing measures, on the plane of dimensions II and III.

laryngitis and, in general, they are low on the third dimension. Two kinds of background noise were also identified: one that is described as crackling, and one that is more like the familiar white random noise. In general, the coders with crackling noise have low values on the third dimension and those with white random noise are in an intermediate position. Thus, the third dimension appears to represent a further distinction between each kind of degradation that could be described as rough vs. smooth. Hoarse speech and crackling noise are rough or irregular in character, muffled speech is smooth or uniform, while speech corrupted by white noise is intermediate between these two extremes.

Objective measures. Figures 3 and 4 also show the vectors corresponding to various objective measures. The vector SNR_{seg} is very close to the coordinate axis of dimension I and is therefore a good indicator of the overall clarity of the processed speech. Log A/R predicts the distribution of points on the second dimension, interpreted as the prevalent kind of degradation, signal distortion or background noise. A low A/R produces a low step size on average, leading to slope overload, perceived as signal distortion. On the other hand, a high value of A/R results in a high average step size and high granular noise. The locations of the vectors $SNRG_{seg}$ and P are also consistent with our interpretation of the coordinate axes. They both have high negative weighting on dimension II because both reflect the predominant impairment category. High P

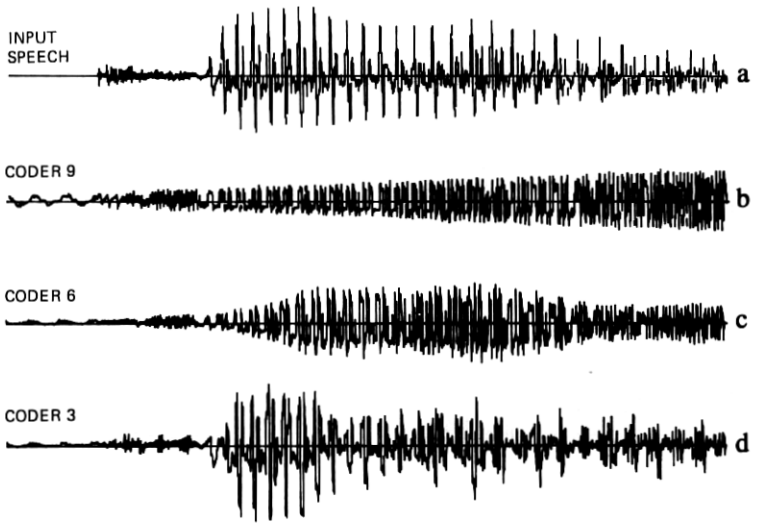

Fig. 5—Waveforms of the word "tool" for 3-bit ADPCM coders low on dimension II (signal distortion), with different values on dimension III, showing the relationship of attack time to the smooth or rough subjective descriptions.

means high overload and substantial distortion; high SNRG_{seg} means low noise. Their non-zero weight on dimension I indicates their influence on speech clarity. A high value of SNRG_{seg} indicates low background noise and enhanced clarity. Conversely, a high value of P is correlated with high distortion and thus with low clarity. The small angle between P and SNRG_{seg} reflects the fact that overload and granularity usually vary reciprocally in coders with a given number of bits per sample.

The coordinate axis of dimension III is highly correlated with $\log T_a$. When the attack time is very high, the step size is very slow in following fluctuations in input level and the speech sounds muffled. When the attack time is very low, the step size frequently overshoots its target value at the beginning of pitch periods, causing irregularity in the periodicity of the processed speech. This irregularity makes the speech sound hoarse.

These properties are apparent in Figs. 5 to 7 which show waveforms that are representative of coder locations in the II–III plane. All of them display the word "tool" processed by 3-bit ADPCM coders with substantially impaired clarity. Figure 5 shows the waveforms of distortion-producing coders (low weighting on dimension II). With −10 dB relative load factor they all produce substantial slope overload in steady state. Coder 9 (Fig. 5b) with $T_a = 32$ msec and high weight on dimension III is the most muffled; fluctuations in the signal envelope are very heavily smoothed. Coder 6 (Fig. 5c), $T_a = 4$ msec, lower on dimension III, reproduces long-term envelope fluctuations, but smoothes out in-

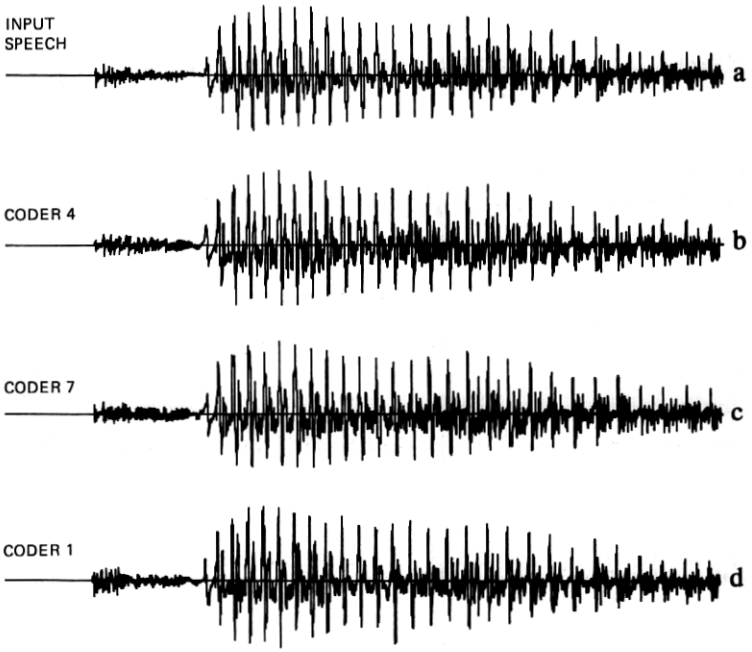

Fig. 6—Waveforms of the word "tool" for 3-bit ADPCM coders high on dimension II (background noise), with different values on dimension III.

dividual pitch periods. Coder 3 (Fig. 5d), T_a = 0.25 msec, moderate loading on dimension III, reproduces pitch contours but with substantial time and amplitude distortion. Figure 6 shows the waveforms of the very noisy coders, 4, 7, and 1, with high relative load factors (4 dB) and high weighting on dimension II. With low distortion they all preserve the general envelope and time structure of the original, so that the nature of their impairment is best seen in oscillograms of noise voltages (coder output minus input). The two extreme types of noise are displayed in Fig. 7. Coder 4, with relatively high weight on dimension III, has "smooth" noise which is shown in Fig. 7b to be correlated with the long term envelope of the speech. In Fig. 7c, coder 1, with crackling noise impairment, low weight on dimension III, is seen to produce impulsive-type noise correlated with the pitch contours of the signal.

Quality prediction. The vector labeled R corresponds to the mean ratings on the 9-point response scale and is very close to the first dimension. As indicated by Fig. 2, a vector corresponding to quality derived from the paired-comparison preference judgments would be in essentially the same location. Multiple regression procedures were used to derive linear relationships between the objective measures and subjective quality. Table III lists the formulas for predicting the ratings from several

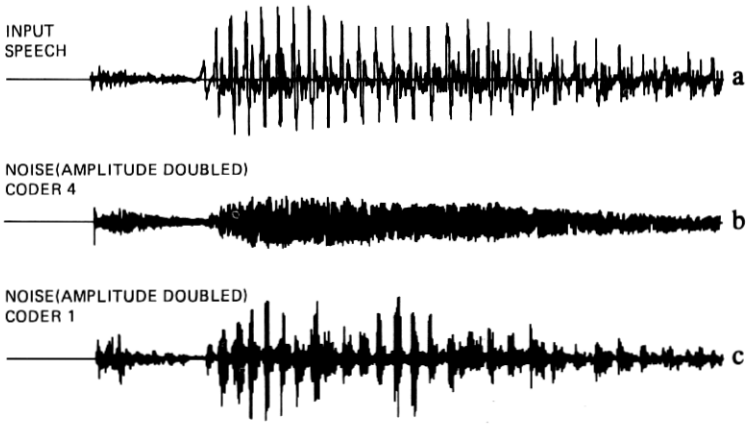

Fig. 7—Noise waveforms relative to the word "tool" processed by two coders of Fig. 6, with the amplitude scale doubled, showing the influence of attack time on the types of noise.

of the objective measures, singly and in combination. Prediction accuracy is indicated by correlations between actual and estimated ratings (1.0 would be perfect agreement), and by the rms error expressed as a fraction of a point on the 9-point scale. Table III shows that, consistent with the locations of their vectors relative to the R vector, SNR_{seg} (formula 2) is a very good predictor of subject quality and SNR (formula 1) is not a good predictor. Among the formulas that contain more than one objective measure, the most accurate predictors of subjective quality are 7 and 8 which include separate measures of granular and overload impairments. Although the location of the vector corresponding to SNRG is essentially the same as that of the SNRG_{seg} vector, prediction accuracy is higher when the measurement is made segmentally.

Other coders. Since formula 8 proved an accurate estimator of the subjective quality of the 18 coders in the experiment, we used it to estimate subjective quality of other ADPCM coders with a wide range of

Table III

	Formula for predicting rating	Corr.	RMS error
1	$0.21\ \text{SNR} + 2.27$	0.69	1.20
2	$0.31\ \text{SNR}_{seg} + 0.89$	0.93	0.63
3	$0.33\ \text{SNR}_{seg} - 0.45 \log A/R + 0.79$	0.95	0.54
4	$0.25\ \text{SNRG} - 0.067\ P - 0.39 \log T_a + 0.22$	0.94	0.55
5	$0.24\ \text{SNRG} - 0.077\ P + 0.68$	0.93	0.63
6	$0.16\ \text{SNRG} - 1.11 \log T_a + 0.47$	0.69	1.21
7	$0.24\ \text{SNRG}_{seg} - 0.078\ P - 0.22 \log T_a + 1.00$	0.96	0.45
8	$0.25\ \text{SNRG}_{seg} - 0.084\ P + 1.19$	0.96	0.48
9	$0.14\ \text{SNRG}_{seg} - 1.03 \log T_a + 1.38$	0.65	1.27
10	$-0.036\ P - 0.26 \log T_a + 4.97$	0.56	1.38

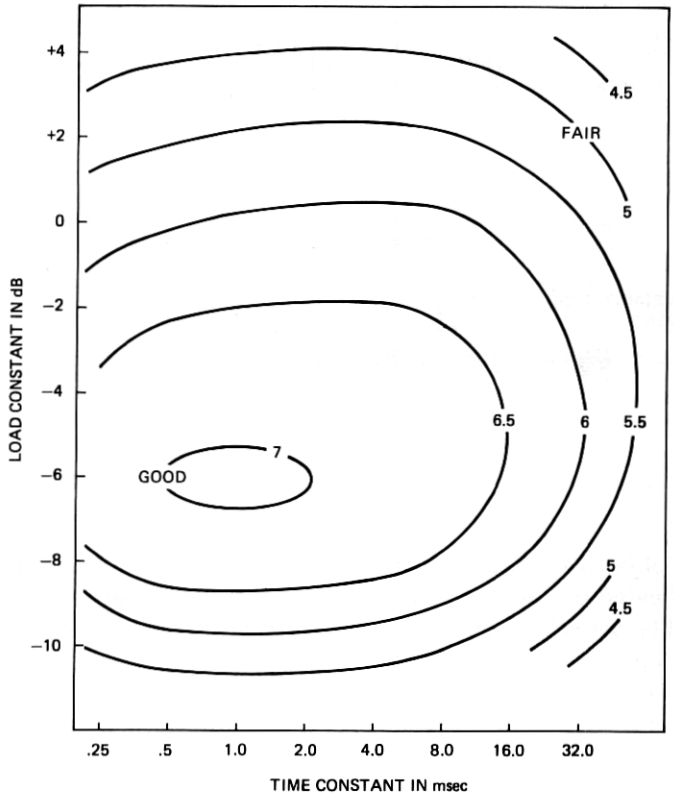

Fig. 8—Equi-rating contours predicted by formula 8, Table III, for 4-bit ADPCM coders.

design parameters. To do so, for each bit rate, we simulated 64 coders that comprised all combinations of 8 load factors and 8 time constants. The 8 values of each parameter included the 3 tested in the original experiment and 5 intermediate values. SNRG_{seg} and P were measured on 4 sentences, one by each talker, processed through each of the 64 coders. The quality ratings, predicted using formula 8 with the averages of the measures on the 4 sentences, are displayed in Figs. 8 and 9, which pertain to 4-bit and 3-bit coders, respectively. The equi-rating contours show that near-optimum quality can be expected over a surprisingly wide range of circuit conditions. For instance, with 4-bit coding, a rating of 6.5 ($\frac{1}{2}$ point from optimum on the 9-point scale) is maintained over a 7 dB range of load factors and a 32:1 range of time constants.

VIII. DISCUSSION

Perceptual characteristics. Our interpretation of the 3-dimensional subjective space is consistent with previous work[16] on analog speech

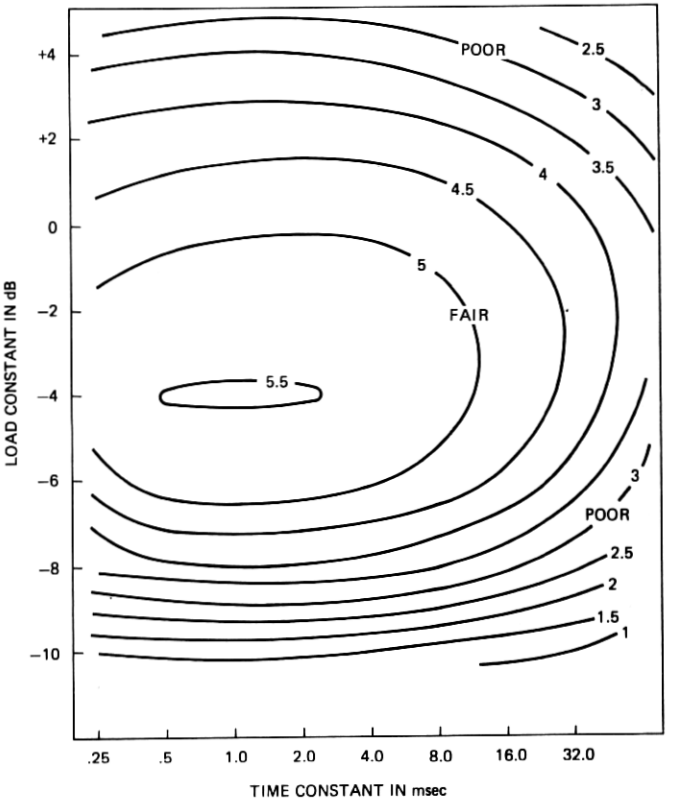

Fig. 9—Equi-rating contours predicted by formula 8, Table III, for 3-bit ADPCM coders.

impairments. In both cases, the first two dimensions have the same meaning. The third dimension in Ref. 16 was related to loudness. In the present experiment, the stimuli were equalized in level, so that subjects could attend to less obvious differences, like the "rough" or "smooth" character of the impairment. The plane of dimensions II and III, Fig. 4, provides perhaps the most interesting view of the subjective space. Accounting for 38 percent of the variance in the average difference judgments, it represents the *kind* of degradation, independent of the *amount* of degradation. In this plane, the perceptually meaningful classifications of ADPCM impairments are the categories, "speech distortion" and "background noise" (dimension II) and, in addition, the types of distortion, "muffled" (smooth) and "hoarse" (rough) and the types of noise, "continuous" (smooth) and "crackling" (rough).

This geometric representation also confirms that the mathematical separation of ADPCM performance into static and dynamic response categories[3] is perceptually meaningful. Dimension II is highly correlated

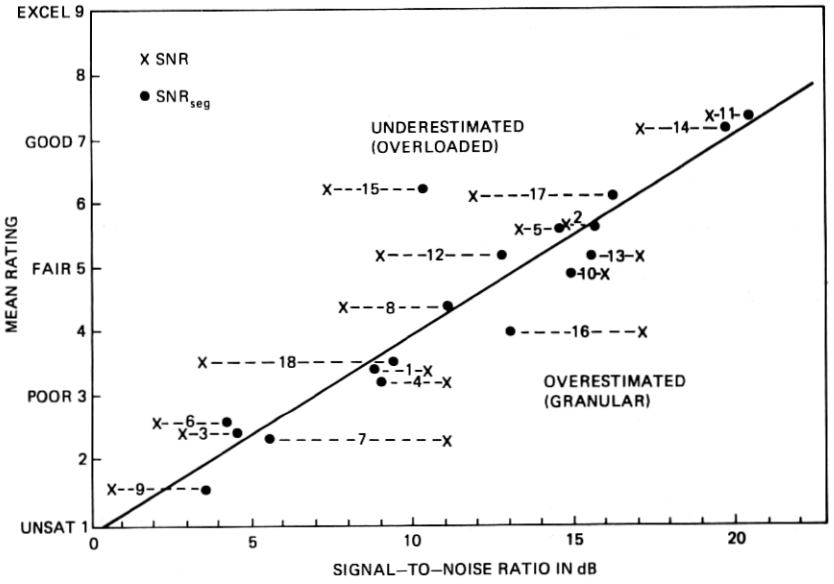

Fig. 10—Mean ratings for the 18 coders vs. SNRs showing the improvement in prediction by measuring SNR segmentally. The solid line is the graph of formula 2, Table III, the regression of rating on SNR$_{seg}$.

with log A/R, a measure of static performance, and dimension III is highly correlated with log T_a, a measure of dynamic performance.

Objective measures. Table III indicates that, as in the case of PCM,[1] ADPCM quality is accurately predicted by a linear combination of overload and granularity measures. Formulas 4, 5, 7, and 8, all containing separate measures of overload and granularity, are among the 6 good predictors of average rating. The table also demonstrates the value of measuring signal-to-noise ratio segmentally. Formula 2, which contains the single measurement, SNR$_{seg}$, is also one of the 6 good predictors. Segmental measures give equal importance to strong and weak components of speech, while non-segmental SNR is essentially a measure of the quality of the high-level components. The strong correlation of SNR$_{seg}$ with average rating indicates that subjective quality judgments are influenced by weak sounds as well as strong sounds.

These properties of SNR and SNR$_{seg}$ are revealed by Fig. 10 which is a scatter plot of average rating vs. both measures for the 18 coders. SNR points are labeled with crosses and SNR$_{seg}$ points are labeled with circles. The line is the graph of formula 2, the regression of R on SNR$_{seg}$. The coders to the left of the line are, for the most part, those with low A/R, in which overload distortion is the predominant impairment. This distortion affects only the strong sounds which are the ones that determine SNR. The good reproduction of weak sounds by overloaded coders is not

reflected by SNR and the crosses for these coders tend to be far to the left of the regression line. That is, they give an unduly poor indication of quality. By contrast, the points to the right of the line tend to be those with high A/R, coders with mainly background noise impairment. This degradation is particularly harmful to weak sounds and therefore its effect is less on SNR than on SNR_{seg}. Consequently, SNR gives an unduly good indication of the quality of these coders.

SNR_{seg} apparently resembles Q, the objective measure of coder quality proposed by D. L. Richards.[17] Q is an average of SNR measures performed with different input levels of a stationary signal. As such it apparently fails to take into account the dynamic response of a coder, which is an important aspect of adaptive quantization. We therefore speculate that as an estimator of subjective quality, the accuracy of Q is intermediate between that of SNR and SNR_{seg}.

Coder design. The relatively large distances between equi-rating contours in Figs. 8 and 9 show that a designer has very substantial latitude in choosing a coder with a prescribed quality rating. This finding is contrary to quality predictions based on conventional SNR measures, which indicate that only restricted sets of design parameters offer near-optimum performance. This newly discovered design flexibility could be valuable in finding coders that simultaneously satisfy criteria in addition to the quality of the coding-decoding process. Examples of such criteria are quality of tandem connections of codecs, resistance to transmission errors, ability to communicate voiceband data, compatibility with other code formats, and economy of implementation.

IX. ACKNOWLEDGMENT

We thank Ann Quinn for recruiting the subjects and running the experiment.

REFERENCES

1. D. J. Goodman, B. J. McDermott, and L. H. Nakatani, "Subjective Evaluation of PCM Coded Speech," B.S.T.J., *55,* No. 8 (October 1976), pp. 1087–1109.
2. D. J. Goodman, J. S. Goodman, and M. Chen, "Intelligibility and Subjective Quality of Digitally Coded Speech," IEEE Trans. on Acoustics, Speech & Signal Processing (in press).
3. D. J. Goodman and A. Gersho, "Theory of an Adaptive Quantizer," IEEE Trans. on Commun., *COM-22,* No. 8 (August 1974), pp. 1037–1045.
4. P. Castellino, G. Modena, L. Nebbia, and C. Scagliola, "Bit Rate Reduction by Automatic Adaption of Quantizer Step Size in DPCM Systems," Int. Zurich Seminar, Zurich, Switzerland, April, 1974.
5. C. Scagliola, "An Adaptive Quantizer with Channel Error Recovery," CSELT Rapporti Technici, *IV,* No. 4 (December 1976), pp. 177–184.
6. D. J. Goodman and R. M. Wilkinson, "A Robust Adaptive Quantizer," IEEE Trans. on Comm., *COM-23,* No. 11 (November 1975), pp. 1362–1365.
7. N. S. Jayant, "Adaptive Quantization with One Word Memory," B.S.T.J., *52,* No. 7 (September 1973), pp. 1119–1144.
8. P. Noll, "Adaptive Quantizing in Speech Coding Systems," Int. Zurich Seminar, Zurich, Switzerland, April, 1974.

9. R. N. Shepard, "Metric Structures in Ordinal Data," J. Math. Psych., *3*, No. 2 (July 1966), pp. 287–315.
10. R. N. Shepard, "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function. I," Psychometrika, *27* (1962), pp. 125–140.
11. R. N. Shepard, "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function. II," Psychometrika, *27* (1962), pp. 219–246.
12. J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika, *29* (March 1964), pp. 1–27.
13. J. B. Kruskal, "Nonmetric Multidimensional Scaling: a Numerical Method," Psychometrika, *29* (June 1964), pp. 115–129.
14. P. Slater, "Analysis of Personal Preferences," Brit. J. Stat. Psychol., *13* (November 1960), pp. 119–135.
15. J. D. Carroll, "Individual Differences and Multidimensional Scaling," in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences—Vol. I: Theory,* Shepard, Romney, Nerlove (eds.), New York: Seminar Press, 1972, pp. 105–155.
16. B. J. McDermott, "Multidimensional Analyses of Circuit Quality Judgments," J. Acoust. Soc. Am., *45*, No. 3 (March 1969), pp. 774–781.
17. D. L. Richards, "Speech Transmission Performance of PCM Systems," Electronics Letters, *1*, No. 2 (April 1965), pp. 40–41.