THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 8, October 1976
Printed in U.S.A.

Subjective Evaluation of PCM Coded Speech

By D. J. GOODMAN, B. J. McDERMOTT, and L. H. NAKATANI

(Manuscript received February 3, 1976)

Subjective quality ratings of PCM *coded speech were obtained with the aims of (i) determining the effects of certain coder parameters and their interactions on speech quality, (ii) finding objective measures for predicting perceived distortions, and (iii) providing guidelines for optimizing coder design. Coders with various combinations of four clipping levels, seven step sizes, four bandwidths, and three logarithmic companding laws were simulated. The coders were rated for quality on a 10-point scale by 48 listeners who heard male and female speech processed by the coders.*

The ratings depended strongly on clipping level and step size, but only weakly on bandwidth. None of the coder parameters interacted strongly with another. Clipping noise power grossly overestimated the extent of perceived overload distortion; instead, clipping percentage is proposed as a much more realistic predictor. Signal-to-granular-noise ratio was a good predictor of perceived granular noise. For a given bit rate, the coder with the highest quality rating was not the coder with minimum total clipping and granular noise power, contrary to traditional wisdom.

I. INTRODUCTION

"How does it sound?" This is a fundamental but elusive question for the engineer designing or evaluating a system for transmitting, recording, or processing speech signals. If the system is analog, the engineer has as a guide a substantial body of information about the interrelated effects on speech quality of such factors as attenuation, noise, linear and nonlinear distortion, echo, and cross-talk.[1] With respect to digital systems, however, the subjective effects of characteristic distortions have been documented to a much smaller extent and, as a consequence, the quality of an existing system and the merits of proposed designs are much harder to predict.

One approach to the evaluaton of digital systems is to relate a digital signal distortion to one of the analog distortions, and to define digital speech quality as the subjective correlate of the equivalent analog distortion.[2] Although expedient and reasonably accurate for certain individual distortions, the value of this approach seems quite limited in the important situations where several distortions occur simultaneously.

While the engineering literature contains many reports of subjective tests of digitally coded speech, most of the tests were undertaken to provide performance data on the overall distortions produced by specific coders. Among the exceptions to this approach and more aligned with the spirit of our work are the experiments reported by Donaldson and Chan,[3] O'Neal and Stroh,[4] and Yan and Donaldson[5] in which individual sources of distortion were identified and the manner of their interaction investigated. In these studies, the effects of bandwidth, predictor network, number of bits per sample and transmission error rate in PCM (pulse code modulation) and differential PCM systems were studied. Another design variable, quantizer overload point, was held fixed although Ref. 5 ends with the suggestion, "A careful study of the dependence of subjective quality on . . . [overload point] . . . seems necessary." Our experiment contains a thorough study of the role of this parameter in PCM.

II. AN OVERVIEW OF THE EXPERIMENT

We used a digital computer to process speech with 208 different PCM coding schemes whose characteristics span an important range of bandwidths, number of bits per sample, overload levels, and compression characteristics. Our aims included the study of: (*i*) the influence on speech quality of the above design parameters, (*ii*) objective measurements that are good predictors of speech quality, and (*iii*) optimum combinations of code parameters.

In the experiment, 48 listeners used a 10-point opinion scale to provide quality ratings of speech processed by each of the coders. The speech material consisted of 10 sentences, each spoken by two females and two males. Our principal conclusions from the analyses of the data are:

(*i*) Overload level and quantizing step size were primary determiners of listeners' ratings. Bandwidth was, by comparison, a secondary determiner of speech quality.

(*ii*) The traditional objective measurement, overall signal-to-noise ratio, was not a useful predictor of speech quality. On the other hand, the percent of samples clipped, P and the signal-to-noise ratio, Q of the granular quantizing noise were useful and independent predictors of speech quality. A simple linear equation

$$R = aP + bQ + c,$$

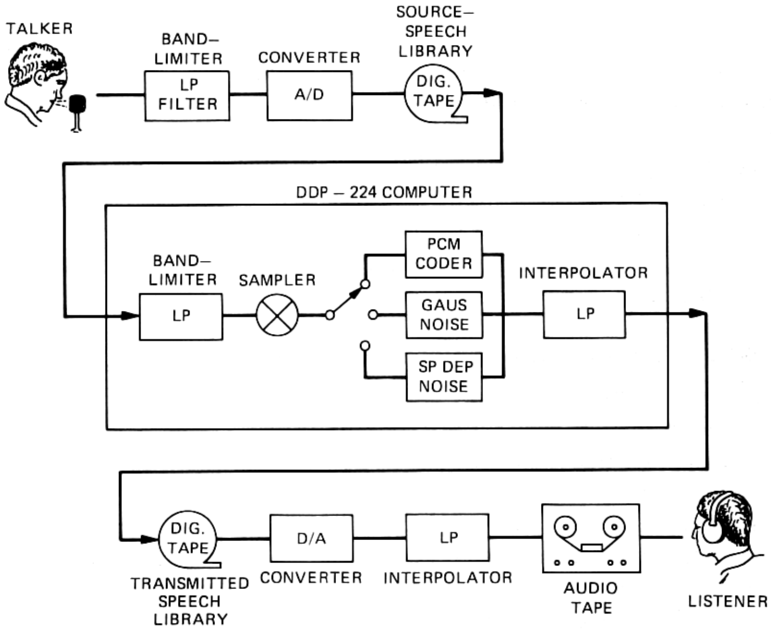


Fig. 1—Block diagram of major steps in the experiment.

where *a, b,* and *c* are empirically derived constants, was a good predictor of the quality rating *R.*

(*iii*) For a fixed number of bits per sample, the coder with the highest quality rating was not the coder with the highest signal-to-noise ratio.

The experiment involved three major steps, as shown in Fig. 1. The first step was to compile a source speech library consisting of high-quality recordings of sentences. The second step was to simulate a variety of coders and noise processes on a computer. The final step was to process the source speech with the simulated coders and noise processes in accordance with an overall experimental design and to obtain subjective quality ratings from listeners.

III. SOURCE-SPEECH LIBRARY

Digital recordings were made of the ten phonetically balanced sentences listed in Table I as spoken by two females and two males. The talkers were seated in a sound-proof booth and spoke into a Sony ECM 22p microphone. The amplified microphone signal was low-pass filtered at 9.6 kHz, sampled 24,000 times per second, uniformly quantized to 12 bits per sample, and written onto digital tape. Each sample was represented by an integer between -2047 and $+2048$. For each talker, the quantizer step size was adjusted manually to use the full quantizer range without clipping. Once a step size was established for a talker, the same step size was used for all ten sentences. This procedure approximately equalized the peak power level of the four talkers over all sentences. The source-speech library thus consisted of digital recordings of 40 sentences containing all the speech sounds spoken by four talkers and approximately equalized for peak power over talkers.

IV. SIMULATION OF CODERS AND NOISE PROCESSES

A PCM system contains a low-pass presampling filter of bandwidth W, a sampler that generates $2W$ equally spaced signal samples per second, a quantizer operating independently on each sample, and a low-pass desampling filter of bandwidth W which generates a continuous waveform from the quantized sequence. In the experiment, each of these components—presampling filter, quantizer, desampling filter—was simulated on a DDP-224 digital computer. Within the computer, "analog speech" appeared in the 24,000-samples/second, 12-bits/sample format of the recording scheme, while sampled and quantized speech appeared with fewer bits and fewer samples.

4.1 Bandlimiting and sampling

The four sampling rates used in the experiment were all integer submultiples of 24 kilosamples/second: 12, 8, 6, and 4.8 kilosamples/second and the nominal cutoff frequencies of the associated low-pass filters were 6, 4, 3, and 2.4 kHz, respectively. The filters, all realized as finite impulse-response digital filters with integer coefficients, were designed to meet the requirements listed in Table II, which conform to

Table I — The ten sentences spoken by each of four talkers *

1. A lathe is a big tool.
2. Grab every dish of sugar.
3. An icy wind raked the beach.
4. Her father failed many tests.
5. Joe brought a young girl.
6. The chairman cast three votes.
7. The boy was mute about his task.
8. Beige woodwork never clashes.
9. Both teams started from zero.
10. My cap is off for the judge.

* Each is a simple declarative sentence that can be spoken in approximately two seconds. The list includes all the phonemes of English in initial, final, and intervocalic position.

Table II — Bandlimiting filter specifications

Sampling rate (kilosamples/s)	12	8	6	4.8
Nominal cutoff W kHz (Attenuation at least 15 dB at $f = W$)	6	4	3	2.4
Passband edge attenuation within ±0.125 dB	4.5	3	2.25	1.8
Stopband edge attenuation at least 30 dB	7.125	4.75	3.562	2.85
Filter order	21	33	41	51
Oversampling ratio	2	3	4	5

the requirements imposed on channel banks of digital multiplex systems.

4.2 Interpolation

The digital interpolating filter simulates the desampling filter of a PCM coder. The latter transforms a sampled signal to a continuous waveform. In the computer, "continuous waveforms" appear as samples occurring at the rate of 24,000 per second; to produce them, an interpolating filter inserts 1, 2, 3, or 4 new samples between each pair of PCM samples, depending on whether the sampling rate of the simulated coder is 12, 8, 6, or 4.8 kHz, respectively.

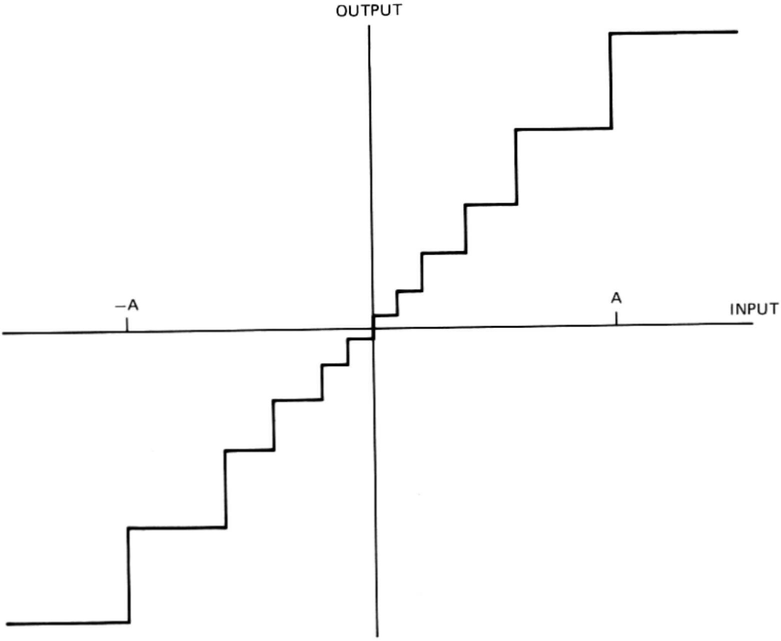


Fig. 2—Input/output diagram of a quantizer.

While, in practice, specifications of desampling filters are often identical to those of presampling filters, we found for our purposes that a 30-dB stopband attenuation was insufficient at certain frequencies. Some speech sounds, particularly nasals, with strong low-frequency components produced audible output tones in the vicinity of the sampling frequency in desampling. For example, the spectrum of sound with considerable energy around 200 Hz has images at 5800 Hz and 6200 Hz when sampled 6000 times per second. Even attenuated 40 dB, these images produced an audible "whistle," which was very distracting to listeners. In the design of interpolating filters, therefore, we specified an attenuation of at least 65 dB near the sampling frequency.

4.3 Quantization

A quantizer is defined by an input/output diagram such as Fig. 2. To study the subjective effects of quantization, it is appropriate to formulate this operation as a sequence of four processes as in Fig. 3: clipping, compression, uniform quantization, and expansion.

Clipping is an inherent part of the quantizing operation. Figure 2

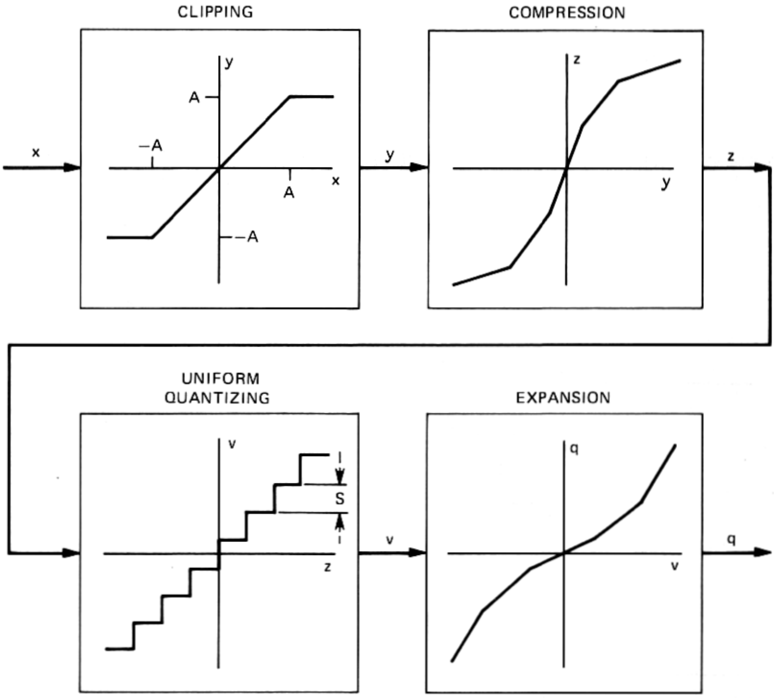


Fig. 3—Four processes included in quantization.

shows that the largest magnitude that can be represented by the quantizer is A. All samples greater than A result in the same output, as do all samples less than $-A$. Hence, the quantizer operates as a device that first clips the input and then represents with finite resolution all signal samples in the range $-A$ to A.

Compression and expansion (each the inverse of the other) are nonlinear transformations of the uniform quantizer's analog input samples and quantized output samples, respectively. Current trends in communication technology favor the use of segment-compression characteristics, which are piecewise linear approximations to logarithmic input/output relationships. In the experiment, we simulated segmented μ-law quantizers* in which the length of each linear segment is double, and the slope one-half, that of the previous segment.[6] The compression curve in Fig. 3 contains five segments. There are three segments for positive inputs and three for negative inputs, with the innermost positive and negative segments colinear. In the input/output characteristic, the quantization step size is constant over a segment and double that of the previous segment. Hence, high-level samples are quantized more coarsely than low-level samples.

In practice, the number of positive (or negative) segments is a power of 2 so that the total number of distinct segments can be written as $2^{(c+1)} - 1$. In the experiment, we studied quantizers with $c = 0$ (uniform quantization), $c = 2$, and $c = 3$, which are 1, 7, and 15 segment quantizers with parameter $\mu = 0$, 15, and 255, respectively.

We describe the uniform quantizer in Fig. 3 by its step size S which is equal to the minimum step size of the nonuniform quantizer of Fig. 2.

The entire quantizer is now defined by three parameters: the overload level A, the companding number μ, and the step size S. For engineering purposes, the most important quantizer parameter is the number of bits per sample B. Table III shows the dependence of B on A, μ, and S over the range of parameters appearing in the experiment. While, in engineering studies, quantizers are usually specified by μ, B, and S or by μ, B, and A, the design and analysis of experiments such as this one are greatly facilitated by viewing μ, A, and S as the independent variables of a quantizer. The advantages of this point of view derive from the fact that quality varies monotonically with both A and S. The relationship of quality to B is considerably more complicated (see Section VIII) and is more readily derived as a combination of two relatively simple functions than measured directly.

Because the source speech appears in the computer encoded in

* The compressor characteristics are piecewise linear approximations to

$$z = sgn(y)[\log (1 + \mu|y|)]/\log (1 + \mu).$$

Table III — Number of bits as a function of step size and clipping level

		$\mu = 0$						
		Clipping level						
		2048	1024	512	256	128	64	32
Step size	1	12*	11	10	9*	8*	7*	6
	2	11*	10	9	8*	7*	6*	5
	4	10*	9	8	7*	6*	5*	4
	8	9*	8	7	6*	5*	4*	3
	16	8*	7	6	5*	4*	3*	2
	32	7*	6	5	4*	3*	2*	1
	64	6*	5	4	3*	2*	1*	
	128	5	4	3	2	1		
	256	4	3	2	1			
	512	3	2	1				
	1024	2	1					
	2048	1						

		$\mu = 15$						
		Clipping level						
		1920	960	480	240	120	60	30
Step size	1	10*	9	8	7*	6*	5*	4
	2	9*	8	7	6*	5*	4*	3
	4	8*	7	6	5*	4*	3*	
	8	7*	6	5	4*	3*		
	16	6*	5	4	3*			
	32	5*	4	3				
	64	4*	3					
	128	3						

		$\mu = 255$			
		Clipping level			
		2040	1020	510	255
Step size	1	7*	6	5	4*
	2	6*	5	4	
	4	5*	4		
	8	4*			

* Indicates quantizers used in experiment.

steps of 1 from −2047 to 2048, A cannot exceed 2048 and S cannot be less than 1. Hence, for each c, there is an upper limit on the number of bits per sample that can be simulated. The limit is 12 bits for $c = 0$, 10 bits for $c = 2$, and 7 bits for $c = 3$. Conversely, there is a lower limit on B because there must be at least one output level for each positive segment and one for each negative segment in the compression curve. This implies that the $c = 2$ quantizer must have at least 3 bits/sample and the $c = 3$ quantizer at least 4 bits.

After a pilot experiment, we decided to vary S in octave steps. Table III shows for each companding law the values of S, A, and B that can be simulated by our procedure. An asterisk indicates a quantizer used in the experiment. The first column of each matrix contains quantizers with no clipping. We omitted quantizers in the second and third column because the pilot study suggested that the deterioration in quality associated with the transition from the first to the fourth column was approximately the same as the deteriorations in the transition from fourth to fifth, and fifth to sixth columns.

4.4 Noise processes

In addition to speech degraded by the PCM coding process, the experiment included speech distorted by two types of additive noise. There were four levels of simulated white gaussian noise added to speech samples; the noise levels were chosen to provide signal-to-noise ratios around 30, 20, 10, and 0 dB. It was felt that gaussian noise is similar in character to the granular quantizing noise of a uniform quantizer ($c = 0$).

In addition, we included four levels of speech-dependent noise.[7] To each sample x_n was added $\pm \rho x_n$, where ρ is the noise-to-signal ratio and the $+$ or $-$ sign is determined by a simulated coin-toss. Thus, the noise magnitude added to each sample is proportional to the magnitude of the sample. This type of distortion is similar in character to the quantizing noise of a companded quantizer in which the noise magnitude increases in a probabilistic sense with signal magnitude. The four speech-dependent noise levels provide the signal-to-noise ratios 30, 20, 10, and 0 dB, where $\text{s/n} = 20 \log (1/\rho)$ dB.

V. SUBJECTIVE EVALUATION OF TRANSMITTED SPEECH

After all simulations were completed, the source speech was processed by the coders and noise processes and the processed speech written onto digital tape to form a transmitted speech library, as shown in Fig. 1. When the library was complete, the transmitted speech was converted from digital to analog and recorded onto audio tape for subjective evaluation.

Four analog tapes were prepared, each containing one example of each of the 240 experimental conditions: (52 coders + 8 noise conditions) × 4 bandwidths. The assignment of talkers to conditions followed a latin square design in a bandwidth by tape-number matrix. Thus, for a given coder or noise condition, a different talker was associated with each of the bandwidths on a single tape. Over the four tapes all 16 talker-bandwidth combinations appeared with each coder and noise condition. For a given bandwidth, each sentence occurred 6 times and each talker 15 times over the 60 noise and coder conditions.

The 240 conditions on each tape were presented in random order to 48 students at a local university, who listened to the stimuli on TDH-39 earphones. Twelve subjects judged the stimuli of each tape. They were asked to "rate each sentence on a scale of 1 to 10 according to its acceptability as a communication link, using 1 to represent the least acceptable, 10 the most acceptable, and the other numbers between 1 and 10 for intermediate ratings." Before the test began, 20 representative conditions were presented to familiarize the listeners with the range of speech quality.

VI. OBJECTIVE MEASURES OF SPEECH QUALITY

Because signal-to-noise ratio (s/n) is the most frequently cited measure of speech quality, the relationship of s/n to subjective appraisal of processed speech is a matter of substantial interest in an experiment such as ours. A very strong inference of our data is that a single s/n statistic—the ratio of the power in the original speech to the power in the difference between processed speech and original speech—is a poor predictor of subjective quality. Instead, we find that the effects of clipping and granular quantization must be considered separately if we are to arrive at a correct prediction of perceived quality. Therefore, we define two noise components: NC, the clipping noise, defined as $y - x$ in Fig. 3, and the granular quantizing noise NG, defined as $q - y$. The total quantizing noise is

$$NQ = q - x = NC + NG.$$

To facilitate measurement of these and other quantities for each quantizer, we produced a digital data tape which, for each utterance passed through each presampling filter, recorded the number of times each possible sample amplitude (from -2047 to 2048) occurred. We used this tape to calculate the power in each filtered utterance, the mean square values of NG, NC, and NQ for each quantizer, and additionally, the percentage of samples clipped by each quantizer. This last statistic, P, proved a better correlate of listener opinions than the mean square value of NC.

VII. RESULTS

7.1 Overview of data analyses

Statistical procedures were applied to evaluate the relative influence of each of the experimental variables on the listeners' ratings. Analyses of variance showed that two variables, clipping level A and step size S, were the major sources of variability influencing the ratings. Multiple regression procedures provided linear estimates of ratings as functions of two objective distortion measures, one related to A, the other to S.

7.2 Determiners of speech quality

7.2.1 Listeners and tapes

An analysis of variance was computed to study the variability of the ratings of the 12 listeners who judged a single tape, and the variability among the four listener groups, each of which judged a different tape. The analysis showed that the listeners within each group were not significantly different in their ratings and that the ratings among the four groups of listeners were not significantly different. In each case, the F ratios were less than 1.0. Therefore, the mean of the listeners' ratings for each condition was used for all further analyses.

7.2.2 *Coder parameters*

A second analysis of variance was computed, using the means of the listeners' ratings, to study the effects of the experimental variables. In this analysis, the differences in the ratings due to step size, clipping, and companding were statistically significant, as expected, and in combination accounted for 84 percent of the total variance. While the variability in the ratings due to the different talkers, the different bandwidths, and their interaction were all statistically significant, each of these effects accounted for only 2 to 3 percent of the total variance.

7.2.3 *Talkers*

Figure 4 shows the mean rating across clipping and step size as a function of bandwidth for each talker at the three companding values

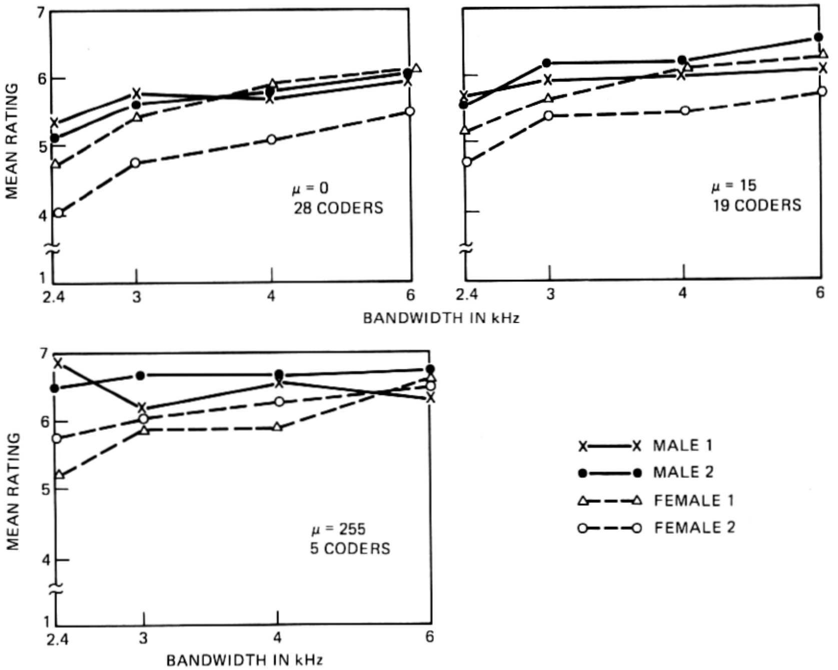


Fig. 4—Mean rating across clipping level and step size as a function of bandwidth for each talker and each companding law.

Table IV — Mean ratings across talkers and listeners

Companding			$\mu = 0$				$\mu = 15$				$\mu = 255$	
Clipping Level			2048	256	128	64	1920	240	120	60	2040	255
6 kHz Bandwidth	Step size	1	9.2	8.8	7.0	4.7	9.1	7.6	6.9	4.7	9.2	5.0
		2	9.2	8.5	6.3	4.4	9.1	7.4	5.3	3.4	7.8	
		4	9.1	8.3	6.1	4.1	8.3	6.2	5.2	3.3	6.5	
		8	8.1	7.1	5.8	4.0	8.7	4.8	3.8		4.6	
		16	8.1	6.3	4.8	3.8	7.6	3.9				
		32	6.3	5.1	4.0	2.4	5.9					
		64	5.0	3.8	3.2	1.3	4.3					
4 kHz Bandwidth	Step size	1	9.0	7.6	6.3	3.6	8.7	7.2	5.3	3.8	7.7	4.8
		2	8.3	7.6	6.4	4.2	8.6	6.7	4.7	3.8	7.3	
		4	8.3	6.8	5.7	4.2	8.2	6.3	5.1	3.2	6.0	
		8	7.2	6.2	4.9	4.3	7.5	5.3	3.7		5.2	
		16	6.7	5.7	4.7	3.6	6.9	3.9				
		32	5.9	4.8	3.6	2.5	5.6					
		64	5.3	3.6	3.1	1.2	4.7					
3 kHz Bandwidth	Step size	1	7.9	6.7	5.3	3.9	8.1	6.4	5.1	3.6	8.2	4.5
		2	7.9	6.6	5.2	3.7	8.7	6.2	4.3	3.4	6.9	
		4	7.2	6.4	4.5	4.1	7.8	6.3	4.4	2.7	5.7	
		8	6.6	5.6	5.3	3.3	6.6	4.6	3.3		5.0	
		16	5.8	4.9	4.4	2.7	5.7	4.1				
		32	4.7	3.9	3.6	2.1	4.5					
		64	3.8	2.9	2.6	1.3	3.6					
2.4 kHz Bandwidth	Step size	1	9.4	7.7	6.0	4.7	8.9	8.0	5.5	3.9	8.4	4.6
		2	9.1	6.9	6.2	4.4	8.4	7.2	5.1	4.0	7.5	
		4	8.2	7.3	6.1	4.3	8.1	6.1	4.1	3.3	6.0	
		8	8.0	6.7	5.6	4.2	8.1	5.7	3.5		5.2	
		16	7.1	5.7	5.1	3.2	6.7	4.3				
		32	5.6	4.5	4.2	3.0	5.8					
		64	4.9	4.4	2.8	1.3	5.0					

$\mu = 0$, 15, 255. Although there was some evidence that the coded speech of female talkers was rated somewhat lower than that of male talkers, the major source of the statistically significant differences among the talkers and the talker-bandwidth interaction was the consistently lower ratings assigned to one female voice. The mean power of her speech was approximately 3 dB greater than that of the other three talkers and the standard deviation of the power about 0.2 dB less, making her speech more sensitive to clipping and filtering. Since the effect of the talkers was minimal, the mean rating across talkers was used for further analyses, thus reducing the variability in the data to that due to the influence of only the physical variables of the coders. The mean ratings across talkers and listeners are shown in Table IV.

7.2.4 Bandwidth

Figure 4 also shows the effect of the four different bandwidths on the ratings. Although the ratings tended to decrease as the bandwidth narrowed, the differences between 6, 4, and 3 kHz were very small. Indeed, the source of the significant differences due to bandwidth was the much lower ratings that resulted from reducing the bandwidth from 3 to 2.4 kHz. The ratings pertaining to three of the talkers contained no significant interactions between bandwidth and the other coder design variables. Only in the data for the female talker with the low ratings were these interactions statistically significant. The most salient of these interactions was between bandwidth and clipping.

7.2.5 Clipping, step size, and companding

Figure 5 shows the mean rating across listeners, talkers, and bandwidth at each step size as a function of clipping level, A, for each of the three companding conditions. The horizontal axes show A decreasing (i.e., the amount of clipping increasing) from left to right.

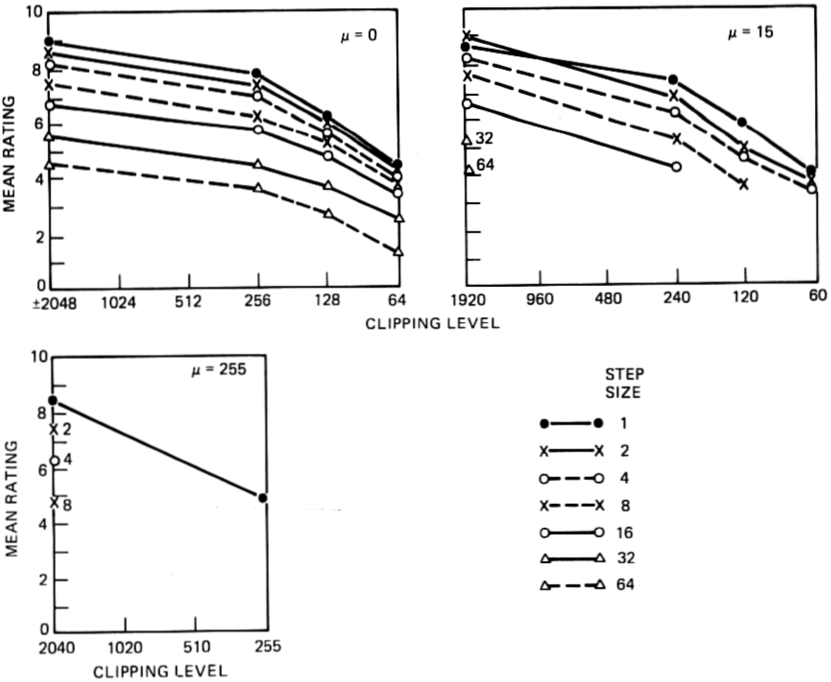


Fig. 5—Mean rating across listeners, talkers, and bandwidth at each step size as a function of clipping level for each companding law.

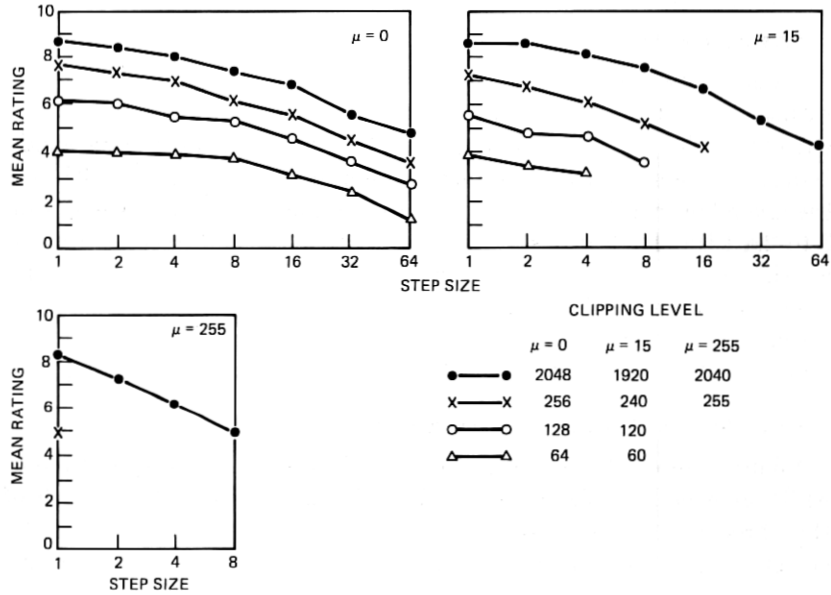


Fig. 6—Mean rating across listeners, talkers, and bandwidth at each clipping level as a function of step size for each companding law.

For each step size S the mean rating decreases as the amount of clipping increases. With $\mu = 0$ or $\mu = 15$, an octave decrease in A results in a relatively small decline in rating when $A > 256$ and a relatively large decrease when $A \leqq 256$. The data for $\mu = 255$, though limited, suggest that relative to the other compression laws, ratings are more sensitive to small amounts of clipping ($A > 256$).

Figure 6 shows the same data points plotted as a function of step size at each clipping level. While, for each A, the ratings are inversely related to S, equal incremental differences in S tend to result in larger differences in the ratings as S increases. That is, the curves generally have a steeper slope when $S > 8$ for $\mu = 0$, and $S > 4$ for $\mu = 15$. The steeper slope of the curve for $\mu = 255$ suggests that quality may be influenced by an interaction between μ and S. An analysis of the ratings of unclipped speech with $S = 1$, 2, 4, and 8 confirms this observation. While the ratings for $\mu = 0$ and $\mu = 15$ were not significantly different, those at $\mu = 255$ were significantly different from the ratings for the other two companding laws.

7.3 Prediction of quality ratings

7.3.1 Signal-to-noise ratio

Figure 7 is a scatter plot of average rating vs measured s/n for the 28 uniform quantizers and the 4 white-gaussian-noise processes. Here, s/n is the usual engineering measure: the ratio of signal power to mean-square difference between quantizer input and output ($x - q$ in Fig. 3). The s/n coordinate of a point in Fig. 7 is the average of the 16 s/n's of the individual utterances (4 bandwidths by 4 talkers) processed by a coder or noise condition. The most important feature of Fig. 7 is the horizontal clustering of the seven points associated with a given value of A, when $A \leqq 256$. In all of these quantizers, the clipping noise, NC, substantially dominates the granular noise, NG, in the total noise, $NC + NG$. This dominance implies that s/n is virtually independent of S with $A \leqq 256$, while, by contrast, perceived distortion depends strongly on S, as evidenced by the vertical spread of the points pertaining to each A. Clearly, in the presence of clipping, s/n is a poor guide to ratings of speech quality: coders with the same s/n elicit widely divergent ratings.

7.3.2 Noise references

In Fig. 7, ratings and s/n are well correlated for one set of coders: those with no clipping, $A = 2048$. Here, the relationship of rating to s/n is similar to that observed for the gaussian noise processes. Figure 7 suggests, therefore, that for uniform quantizing, white gaussian noise is a good noise reference when there is no clipping; it is a poor noise reference when clipping is significant.

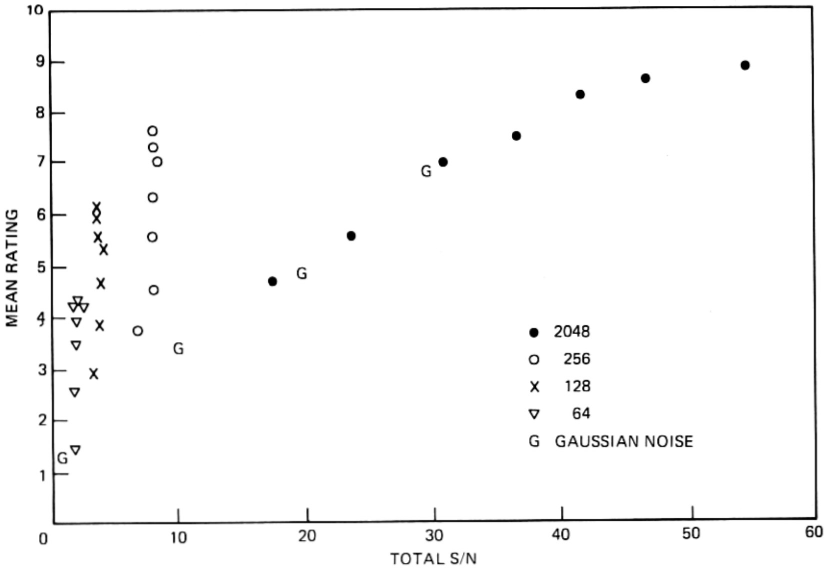


Fig. 7—Relationship between mean rating and total s/n for the 28 linear conditions and the 4 gaussian-noise conditions.

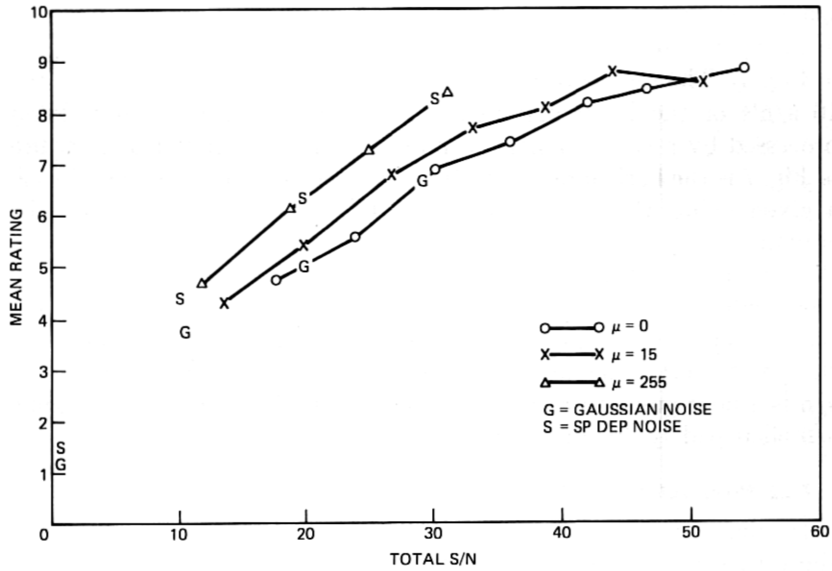


Fig. 8—Mean ratings of unclipped speech for each companding law and the added noise conditions as a function of the total s/n.

Because the amount of granular distortion produced by a coder with companding depends on signal amplitude, one may suppose that speech-dependent noise would provide a better noise reference than white gaussian noise for companded coders.[2,7] Figure 8 lends support to this conjecture by showing, for all of the coders with no clipping and all of the noise processes, the relationship of average rating to s/n. For $\mu = 255$, the relationship is similar to that observed with speech-dependent noise; for $\mu = 0$, it is similar to that observed with white gaussian noise. $\mu = 15$ is an intermediate case.

7.3.3 Regression analysis

Because total s/n proved a poor correlate of listener ratings, we turned to multiple regression procedures to find an objective predictor of the ratings. The analyses of variance indicated that the ratings were primarily influenced by S and A, which have nearly independent effects.* Consequently, we used as independent variables of the regression one distortion measure related to S and one related to A. Appropriate measures proved to be Q, the granular s/n (ratio of signal power to power in NG) measured in dB, and P, the clipping probability (the proportion of speech samples $> A$), expressed as a percentage.

* In the analyses of the linear conditions, S accounted for 35 percent of the total variance, A for 45 percent, and the interaction of S and A for only 1 percent.

The regression was computed at each companding level with the original values for each talker at each bandwidth included as repeated observations. For example, each of the 28 coders with $\mu = 0$ was represented by 16 measurements (4 talkers × 4 bandwidths). Table V lists the coefficients obtained by the regression procedure at each companding level and also by combining the three companding levels. While the values of the coefficients change with μ, the correlations and the rms values of the residuals do not change radically. The regression procedure was applied to the data for the coders with $\mu = 255$ for completeness, but the computation was based on information for only five coders at the smaller step sizes and only one clipping condition. When the ratings of the three companding laws were also included as repeated observations, the ratings predicted by appropriate weighting of only the Q and P correlate highly with the obtained ratings.

VIII. DISCUSSION

8.1 Effects of design variables

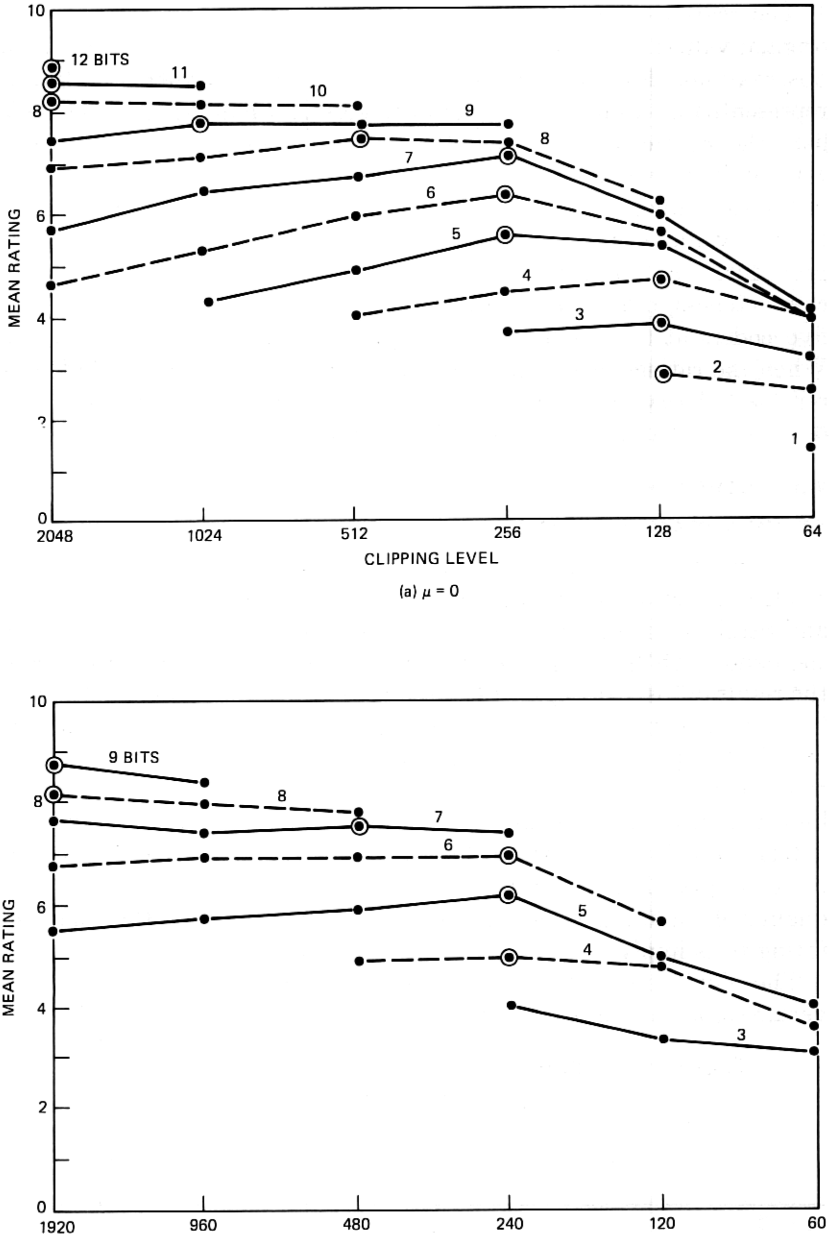
Among the PCM design variables, system bandwidth W had the smallest effect on the ratings, a finding consistent with that of O'Neal and Stroh[4] who state that "over the range of 2.4–4.3 kHz changes in the bandwidth W of the speech signal are inconsequential in terms of the resultant user ratings." (To describe our data, we would substitute "of minor importance" for "inconsequential.") In considering the practical application of this conclusion, a caveat is necessary. In a recent experiment, Goodman, Goodman, and Chen[8] found that band-limiting, although less important than clipping and quantizing in determining listener ratings, had a very strong effect on consonant intelligibility. This suggests that the impact of band-limiting on the quality of communication may be more substantial than the results of rating tests imply.

The significance of the dependence of ratings on the quantizer variables, μ (compression law), S (step size), and A (clipping level),

Table V — Coefficients obtained by regression level
$(\hat{R} = aP + bQ + c)$

	a	b	c	Correlations	RMS Residual
$\mu = 0$	−0.08	0.09	3.87	0.87	1.005
$\mu = 15$	−0.11	0.11	3.09	0.87	0.957
$\mu = 255$	−0.27	0.16	2.96	0.84	0.825
Combined	−0.10	0.09	3.99	0.85	1.038

P = Percent clipped.
Q = s/n granular quantizing noise.



(a) $\mu = 0$

(b) $\mu = 15$

Fig. 9—Mean ratings at a constant number of bits as a function of clipping level for (a) $\mu = 0$ and (b) $\mu = 15$. A circle indicates the highest-rated quantizer of a certain number of bits.

will be more apparent if we return to the usual engineering description of a quantizer which includes the number of bits per sample, B, as an independent variable. Thus, in Fig. 9, we have plotted the same points that appear in Fig. 5, but in this case, we have drawn lines showing contours of constant B rather than constant S. Here we see the effect on subjective ratings of the well-known compromise between clipping and quantizing in coder design. At the left of each curve, we have the quantizers that cause little or no overload but have high step sizes and, therefore, substantial granular quantizing noise. At the right, clipping distortion predominates over granular noise.

Figure 10 demonstrates the effect of companding on ratings by displaying on the same graph rating vs clipping level curves for 5-bit and 6-bit encoders with $\mu = 0$ and $\mu = 15$. For a given clipping level, even this small amount of companding (practical values of μ are 100 and 255) produces substantially higher ratings than those given the uniform quantizer. The companding advantage is well known and accounts for the presence of compandors in all PCM transmission systems. In terms of statistical signal theory, we may explain the advantage by saying that a nonuniform quantizer provides a better match to the probability distribution of speech amplitudes than a uniform quantizer. A perceptual explanation is that the low-level portions of a speech signal carry the most information. With A and B

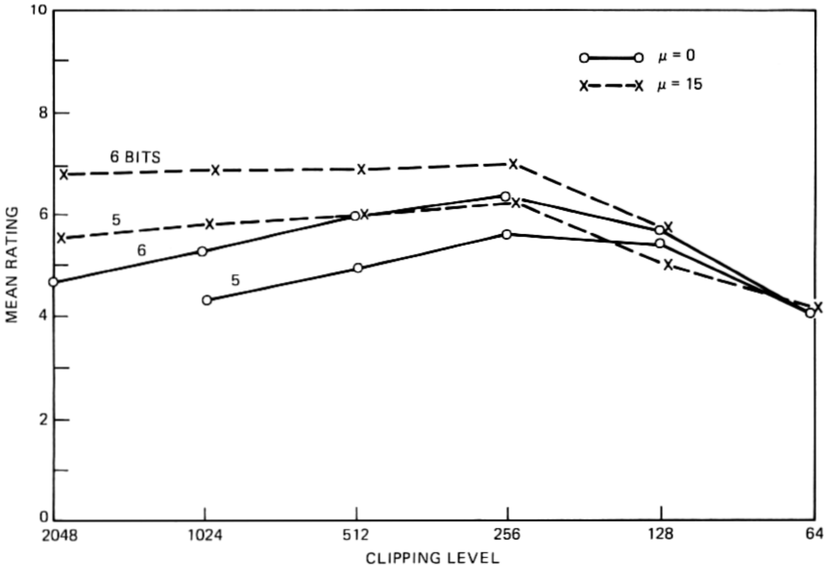


Fig. 10—Mean rating as a function of clipping level for 5 and 6 bits at $\mu = 0$ and $\mu = 15$, demonstrating the effect of companding.

given, a nonuniform quantizer codes low-amplitude samples with a smaller step size (and, therefore, lower NG) than the corresponding uniform quantizer.

8.2 Objective measures of distortion

Figure 11 shows, for coders with uniform quantizers, total s/n as a function of A. There are striking differences between these curves and Fig. 9. The most important differences are in the locations of the maximum points on corresponding curves and the substantially steeper slopes to the right of the maxima in Fig. 11. Both of these differences reflect the fact that NC increases very rapidly from zero as the clipping level decreases from $A = 2048$, while, by contrast, listener opinions are relatively insensitive to clipping until $A < 512$.

The disparity between Figs. 9 and 11 suggests that even with B constant, s/n, the usual engineering measure of quantizer quality, is a poor guide to subjective ratings, mainly because the mean-square clipping is a poor predictor of listener ratings. A more useful measure of clipping distortion is clipping probability, which we have measured as the percentage of samples clipped in an utterance. P varies with A in the manner shown in Fig. 12. Observe that, like the ratings, P changes slowly as A decreases from 2048 and that it is most sensitive to changes in A when $A < 512$. These similarities account for the accuracy of the regression formulas in Table V, which have as in-

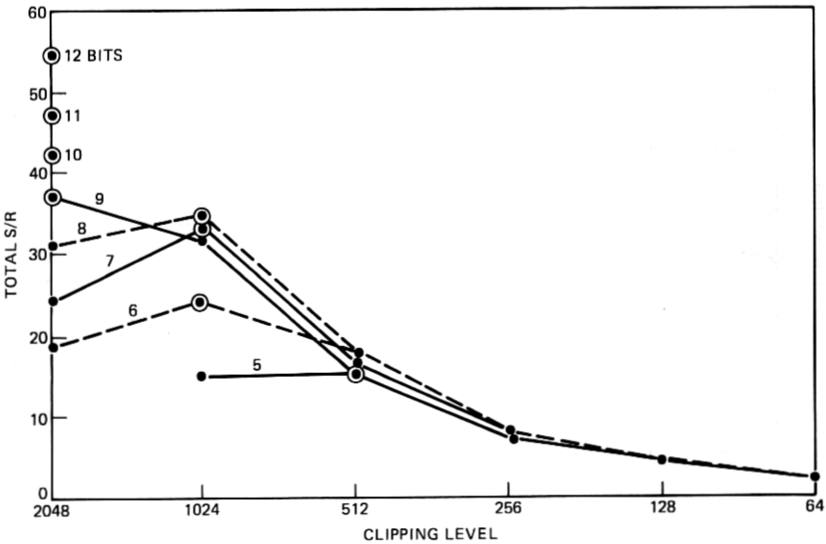


Fig. 11—Total s/n as a function of clipping level at a constant number of bits. Circles indicate the maximum s/n at each bit rate.

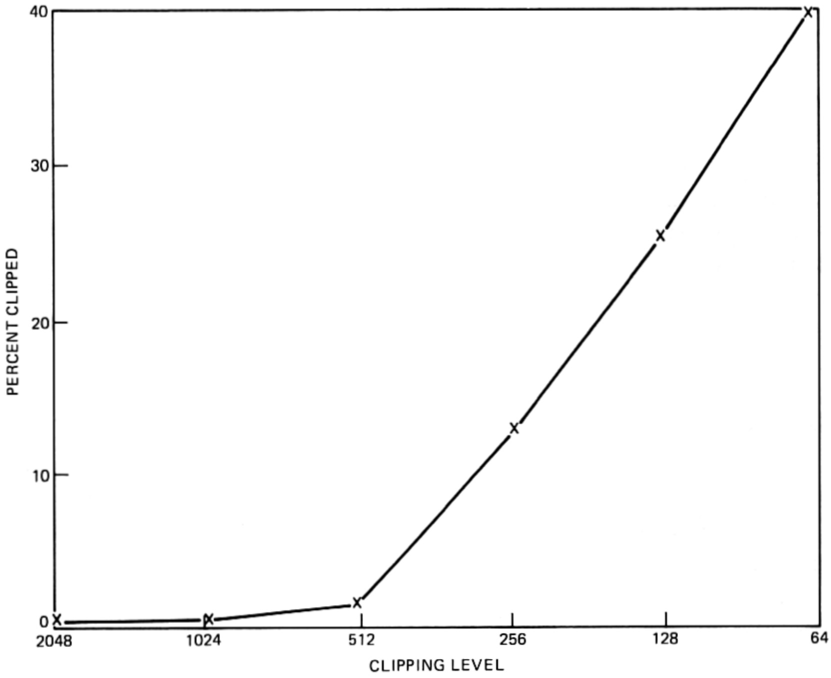


Fig. 12—Percent clipped as a function of clipping level.

dependent variables P and the signal-to-noise ratio Q of the granular quantizing noise NG. These formulas, based on the notion of two perceptually distinct distortions, are more useful predictors of subjective quality than is total s/n, which is based on the fallacious assumption that listeners attend only to the difference between quantizer input and output with no regard to the components of this difference.

8.3 Optimum quantizers

Given a companding law and a fixed number of bits per sample, there is an optimum quantizer with overload point A^* that provides the best mixture of clipping distortion and granular noise. For $A < A^*$, clipping is the predominant type of distortion; for $A > A^*$ granular noise predominates. A circle in Fig. 9 indicates the subjectively optimum overload point for a given bit rate. As the number of bits per sample increases, so does A^*. In high-resolution quantizers, it is possible to have low granular noise and very little clipping simultaneously. Notice that the optimum points in Fig. 9 are all one or two octaves to the right of the corresponding points in Fig. 11. The experiment demonstrates that listeners are more tolerant of clipping than s/n measurements suggest. In addition, the curves in Fig. 9 are considerably broader than those in Fig. 11, which indicates that listeners are relatively insensitive to changes in A in the region of A^*.

This observation relates directly to the quantizer dynamic range problem. While in the experiment we have held the speech power fixed and varied A, we would have obtained the same distortions by holding A fixed and changing the speech signal level. It follows that the horizontal axes that we have labeled "clipping level" can, for a single quantizer, be renamed "speech level," which increases from left to right. Figure 11 shows that a uniform quantizer has near-optimum signal-to-noise performance for only a narrow range of speech levels. By contrast, we see in Fig. 9 that listeners give nearly optimum ratings over a much wider range of input powers.

IX. CONCLUSIONS

Our results lead to several general observations regarding the subjective evaluation of speech degraded by digital coding. First, our data indicate that when the degraded speech includes certain types of digital signal distortions, such as peak clipping, then total s/n is a poor objective indicator of subjective speech quality. For the coders we studied, a simple linear combination of two objective measures was a good predictor of the subjective quality of speech with quantizing and clipping distortions; however, we do not know of any single objective measure which would be a good composite indicator of subjective speech quality for all types and combinations of digital signal distortions. Second, because some types of digital signal distortions seem to be perceptually distinct, it seems unlikely that the subjective quality of digital speech can be evaluated by reference to a single type of analog or digital signal distortion, such as speech-dependent noise. And third, because coders are optimized by trading off different types of distortions, it follows that the important cases to study are those where distortions occur in combination rather than singly. This implies that knowing the relationships between subjective speech quality and various types of reference-signal distortions occurring singly—be they digital or analog—may be of limited value for predicting the subjective quality of coded speech if most practical coders produce speech degraded by combinations of distortions. These observations should be kept in mind by designers who must struggle with the problem of how various parameters of their coders affect the subjective quality of the speech.

REFERENCES

1. D. L. Richards, *Telecommunication by Speech*, London: Butterworths, 1973, Ch. 4.
2. H. B. Law and R. A. Seymour, "A Reference Distortion System Using Modulated

Noise," Proc. Inst. Elec. Eng. London, Part-B, Electron., *109*, No. 48 (November 1962), pp. 484–487.

3. R. W. Donaldson and D. Chan, "Analysis and Subjective Evaluation of Differential Pulse-Code Modulation Voice Communication Systems," *IEEE Trans. Commun. Technol., COM-17*, No. 1 (February 1969), pp. 10–19.
4. J. B. O'Neal, Jr. and R. W. Stroh, "A Speech Encoder-Multiplexer Feasibility Study," Air Force Office of Scientific Research, Report No. AD 739965, March 8, 1972, Ch. 4.
5. J. Yan and R. W. Donaldson, "Subjective Effects of Channel Transmission Errors on PCM and DPCM Voice Communication Systems," *IEEE Trans. Commun., COM-20*, No. 3 (June 1972), pp. 281–290.
6. H. Kaneko, "A Unified Formulation of Segment Companding Laws and Synthesis of Codecs and Digital Compandors," B.S.T.J., *49*, No. 7 (September 1970), pp. 1555–1588.
7. M. R. Schroeder, "Reference Signal for Signal Quality Studies," *J. Acoust. Soc. Amer., 44* (1968), pp. 1735–1736.
8. D. J. Goodman, J. S. Goodman, and M. Chen, "Relationship of Intelligibility and Subjective Quality in Digitally Coded Speech," Talk given at the 90th Meeting of the Acoustical Society of America, Abstract in *J. Acoust. Soc. Amer., 58*, Supplement No. 1 (Fall 1975), p. S130.