THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 3, March 1979
Printed in U.S.A.

A Comparison of the Performance of Four Low-Bit-Rate Speech Waveform Coders

By J. M. TRIBOLET, P. NOLL, B. J. McDERMOTT, and R. E. CROCHIERE

(Manuscript received May 23, 1978)

Subjective ratings were obtained for four different speech waveform coders of varying complexity at each of three bit rates (24, 16, and 9.6 kb/s). ATC (adaptive transform coding) is rated the highest and ADPCM-F (adaptive differential PCM with a fixed predictor) the lowest, regardless of bit rate. Although there is large variability in the ratings due to different talkers and listeners, SBC (sub-band coding) is rated higher than ADPCM-F and about equal to ADPCM-V (adaptive differential PCM with a variable predictor) for most talkers and listeners. A weighted combination of two objective measures, one accounting for noise and the other for bandwidth effects, appears promising as a predictor of the subjective quality ratings.

I. INTRODUCTION

The quality of the reproduced speech from waveform coders can usually be improved by increasing their complexity. However, increasing the complexity also usually increases the cost. The practical problem in choosing a coder, at the current state of the art, becomes one of knowing how much loss in quality is sacrificed when opting for a less complex (less expensive) coder.

This paper compares subjective quality ratings for four different speech waveform coder algorithms of varying complexity. The algorithms, rated in order of their complexity are: ADPCM-F (adaptive differential pulse code modulation with a fixed predictor), SBC (sub-band coding), ADPCM-V (adaptive differential PCM with a variable predictor), and ATC (adaptive transform coding). Each of these four algorithms was studied at three different transmission rates: 24, 16, and 9.6 kb/s. These coders were chosen because they represent a number of different classes of coding techniques ranging from relatively simple (inexpensive) to highly complex (costly) schemes. The choice

was also influenced by the availability of coder software at the time of the experiment. All coders are "non-pitch-predicting."

Since the coders being studied produce a broad range of qualities and types of degradations, the data also provided an opportunity to evaluate the relative merits of several objective measures that have been proposed for predicting quality ratings.[1]

II. THE WAVEFORM CODERS: A BRIEF DESCRIPTION

The four waveform coders used in this experiment are depicted in Figs. 1 through 4 and are briefly described below.

2.1 ADPCM with a fixed predictor (ADPCM-F)

The ADPCM-F coder is the simplest of the four coding techniques. As seen in Fig. 1, it consists of a quantizer with an adaptive step-size and a first-order fixed predictor. The step-size adaptation is based on the one-word memory approach of Jayant, Flanagan, and Cummiskey.[2] The number of bits used in the quantizers are 3, 2, and 1 bit, respectively, corresponding to the transmission rates of 24, 16, and 9.6 kb/s. In the case of the 9.6-kb/s transmission rate, the coder reduces to that of an adaptive delta modulator with a sampling rate of 9.6 kHz. The output of the 9.6-kb/s coder was filtered with a 0 to 2800-Hz lowpass filter to remove the high frequency noise. The parameters used for the ADPCM-F coders are close to those proposed by Jayant.[3]

2.2 Sub-band coding (SBC)

In the sub-band coder, the speech band is partitioned into sub-bands.[4,5] Each sub-band is effectively lowpass-translated and sampled at its Nyquist rate. It is then preferentially encoded using APCM

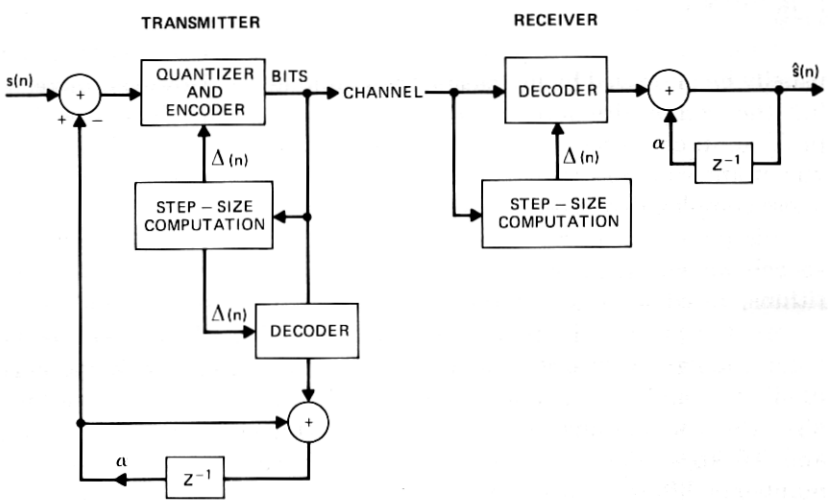


Fig. 1—Block diagram of ADPCM-F.

(adaptive step-size PCM) encoding with a backward step-size adaptation algorithm. The number of bits per sample in each band is chosen according to perceptual criteria for that band. On reconstruction, the sub-bands are decoded and bandpass-translated back to their original bands. They are then summed to give a replica of the original speech signal.

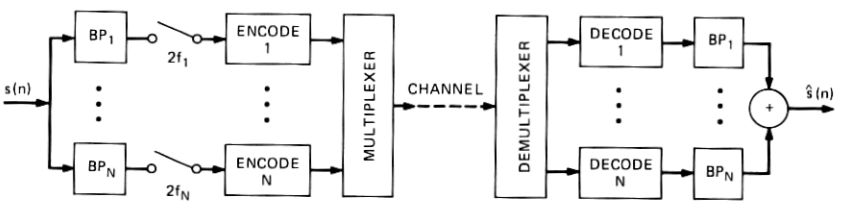
Figure 2a shows an implementation of the sub-band coder based on an integer-band sampling technique.[4] The speech band is partitioned into N sub-bands by bandpass filters BP_1 to BP_N. Typically, four to five bands are used and, at 9.6 kb/s, gaps are permitted between the bands to conserve bandwidth and, therefore, bit rate, as is illustrated in Fig. 2b.

The complexity of the sub-band coder is somewhat greater than that of the ADPCM-F coder. Using recent CCD (charge-coupled device) technology, the filters can potentially be implemented very efficiently. A single APCM coder can be multiplexed between the N sub-bands. The multiplexer requires digital logic and a ROM (read only memory) for storing the multiplexing pattern.

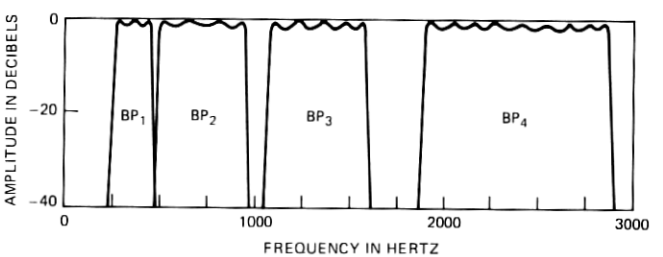
The coder configurations at 9.6 and 16 kb/s are those of examples A and D in Ref. 5. The 24-kb/s coder uses the same filters as the 16-kb/s coder, but the number of bits per sample per sub-band is increased by one.

2.3 ADPCM with a variable predictor (ADPCM-V)

The ADPCM-V coder is a more sophisticated version of ADPCM,[6,7] as seen in Fig. 3. The input data are first buffered and delayed. From this



(a)



(b)

Fig. 2—(a) Block diagram of SBC. (b) Sub-band partitioning.

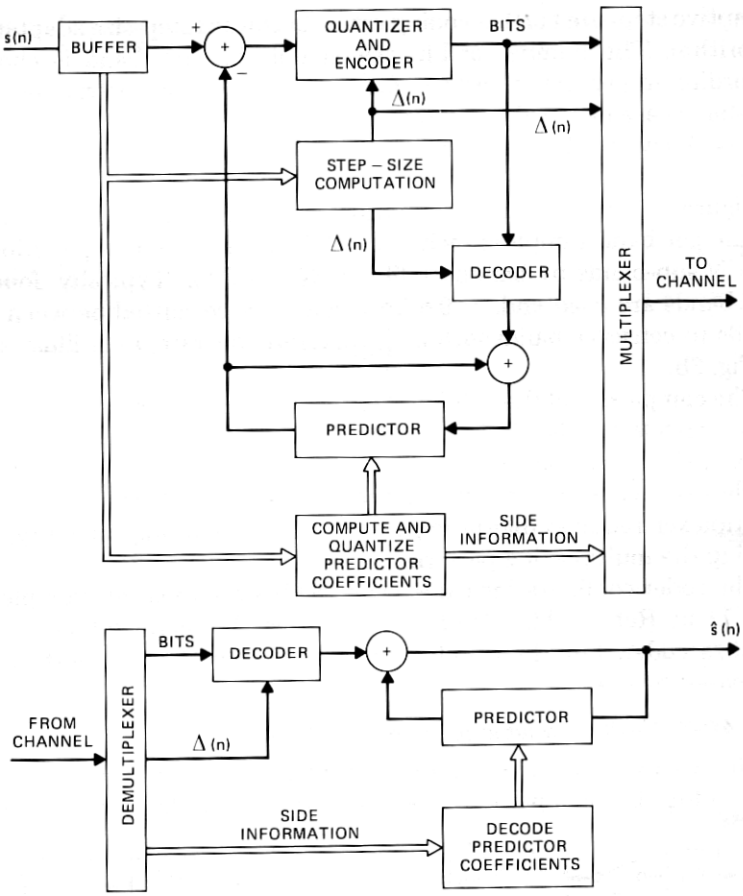


Fig. 3—Block diagram of ADPCM-V.

buffered block of speech, a short-time estimate of the variance of the input speech is computed and used to control the step-size of the quantizer. This local variance estimate is also quantized and transmitted for use in the receiver.

The predictor is an eighth-order adaptive predictor (no pitch prediction is involved in this scheme). The coefficients of the predictor are computed according to the relation

$$\mathbf{h}_N = \mathbf{R}_N^{-1}, \mathbf{r}_N, \tag{1}$$

where $\mathbf{R}_N$ and $\mathbf{r}_N$ are the matrix and vector of autocorrelation coefficients of the data in the buffer. The predictor coefficients are also quantized and transmitted to the receiver as side information. The total transmission of side information requires about 2 kb/s of data.

In the implementation of the ADPCM-V coder, 3, 2, and 1 bits/sample were used for the respective transmission rates. A sampling rate of

8 kHz was used. Since approximately 1.5 to 2 kb/s of additional information are required to be transmitted with this scheme, the actual transmission rates represented by this coder were 26, 18, and 9.6 kb/s instead of 24, 16, and 9.6 kb/s, which was used in the other coders.

The complexity of the ADPCM-V coder is primarily dominated by the implementation of the adaptive predictor and the computation of the predictor coefficients. As seen in eq. (1), this involves an autocorrelation computation and the solution of 8 simultaneous linear equations every 8 to 16 ms. This must be done using high-speed digital computation. Therefore, the complexity of the ADPCM-V coder is substantially greater than that of the ADPCM-V or sub-band coders.

2.4 Adaptive transform coding (ATC)

The adaptive transform coder is analogous in some respects to the sub-band coder in that it divides the speech band into a number of frequency components.[8] The resolution (number of bands), however, is generally much finer than that used in the sub-band coder, and the translation to the frequency domain is achieved by means of a fast transform algorithm. The transform is a 128-point, discrete, cosine transform. The transformed coefficients are encoded with APCM encoding.

Figure 4 is a block diagram of the transform coder. The input speech

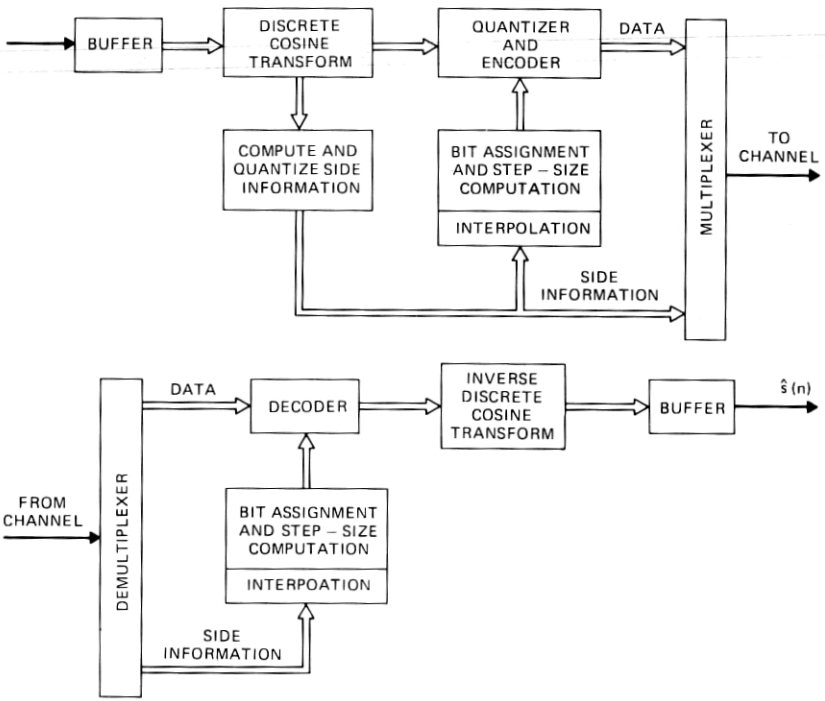


Fig. 4—Block diagram of ATC.

is buffered and transformed in blocks of data. The output values of the transform are smoothed and decimated to 16 values. These 16 values are quantized and sent as side information to the coder and the decoder, where they are interpolated, yielding a smoother version of the short-time spectrum. This smoothed spectrum forms the basis for a dynamic bit assignment and step-size adaptation of the quantizers, as a function of frequency. The quantizers are then used to encode the transform coefficients.

In the receiver, a similar bit assignment and step-size computation is performed and the transform coefficients are decoded. The decoded values are then inverse-transformed to give a replica of the input speech.

The degree of complexity of the transform coder is of approximately the same order as the ADPCM-V coder. The side information and bit assignment computation require high-speed digital computations. Some efficiency might be gained using CCD technology to implement the transform.

III. SUBJECTIVE EVALUATION

3.1 Experimental procedure

Digital recordings of sentences spoken by four talkers (two male and two female) were processed by each of the 12 coders. The processed utterances were equalized to the same mean power to eliminate loudness differences. Two analog test tapes were prepared that contained different permutations of four random orderings of the 12 coders. The talkers were assigned in a balanced design so that each coder was represented by the speech of a different talker in each of the random orders. Since each of the four talkers had recorded a unique set of eight sentences, the sentences were randomly assigned and none occurred more than twice.

Students from the junior and senior classes of local high schools served as paid subjects. They listened to the processed speech binaurally over Pioneer SE700 earphones while seated in a double-walled sound booth. Sixty-five subjects judged the 48 coded sentences (4 coders $\times$ 3 bit rates $\times$ 4 talkers). They were asked to rate the quality of each sentence on a scale from 1 to 9, using a 1 to represent the worst quality, 9 to represent the best quality, and the numbers between 1 and 9 for intermediate evaluations. Before the test session began, they judged six representative conditions for practice to familiarize them with the task and the range of quality.

3.2 Results

An initial analysis of the ratings by the 65 listeners revealed a large amount of unexpected variability in the data due to the different talkers and listeners.

The variability due to the different listeners is illustrated in Fig. 5. These histograms show the percent of the listeners who assigned each of the nine possible ratings to each of the 12 coders. While these values were computed by summing across the four talkers, comparable plots for the individual talkers produced essentially the same general pattern, indicating that the variability in the ratings cannot be attributed entirely to the effects of the different talkers. The extremely skewed distributions for ATC at 24 kb/s and ADPCM-F at 9.6 kb/s show strong agreement among the listeners about which coders had the best and worst qualities, but the wide range, and in some cases almost flat character, of other distributions indicate that the listeners differ in the

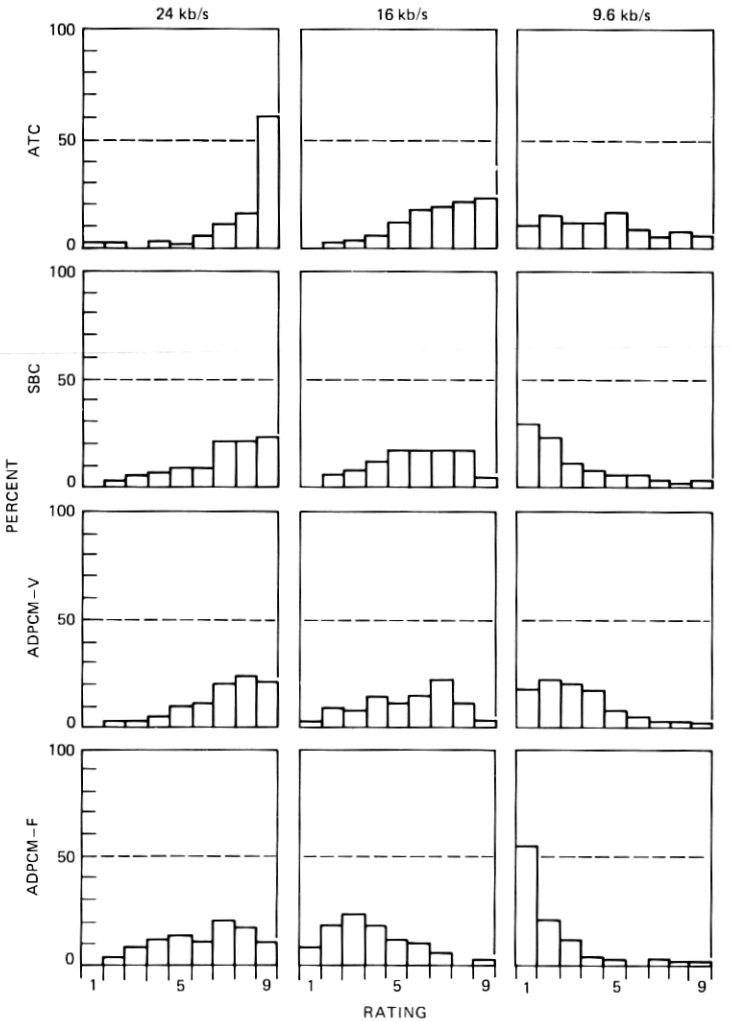


Fig. 5—Percent of ratings assigned each coder (65 listeners × 4 talkers).

amount and type of distortion that they will tolerate. However, it is not the purpose of this paper to examine the sources of listener variability but to gain some information about general trends and the trade-off between complexity and quality. Since so many of the distributions showed a lack of unimodality, the median ratings were used to compare the evaluations according to the experimental variables: type of coder, bit rate, and talkers.

The variability in the ratings due to the different talkers is shown in Fig. 6, where the median rating (bracketed by their 0.95 confidence interval) of the 65 listeners for the 12 coders are plotted for each of the four talkers. The overall pattern for all talkers shows that ATC had the highest rated quality, ADPCM-F the lowest rated quality, and the rating for each talker-coder combination drops as the bit rate is reduced. The ratings for SBC and ADPCM-V are about the same for all voices except female 2, indicating that they have about the same overall quality. For the same three voices, ATC at 16 kb/s is rated equal to, or better than, the other three coders at 24 kb/s.

Some specific differences in ratings are consistent with the usual male-female voice distinctions, and some are speaker idiosyncratic. ATC at 16 kb/s is rated almost equal to 24 kb/s for male voices but much lower for female voices. ADPCM-V at 16 kb/s is rated almost equal to 24 kb/s for female 1 but almost three categories lower for female 2. The coders span a much wider range of ratings at the lower bit rates for male 2 than for male 1. The separation of the four coders into only two classes for female voice 2 is a unique result in this small sample of four voices, but an analysis of the characteristics of her voice could provide important information about the effect of different voices in digital coding techniques.

IV. OBJECTIVE MEASURES

Several objective measures that may be more sensitive to the types of degradations produced by waveform coders are described in Ref. 1. The data from this experiment provided an opportunity to test these theoretically derived predictors of quality ratings. The median ratings (bracketed by the 0.95 confidence interval) computed across both listeners and talkers are plotted in Fig. 7 and served as a single basis for comparing the efficacy of these various objective measures.

4.1 Segmental S/N ratio

Perhaps the most widely used measure of performance has been the conventional signal-to-noise ratio, although it has not generally been found to be a very good indicator of subjective quality. An improved measure, proposed by Noll,[1,9] averages the s/n ratio in short (20 to 30

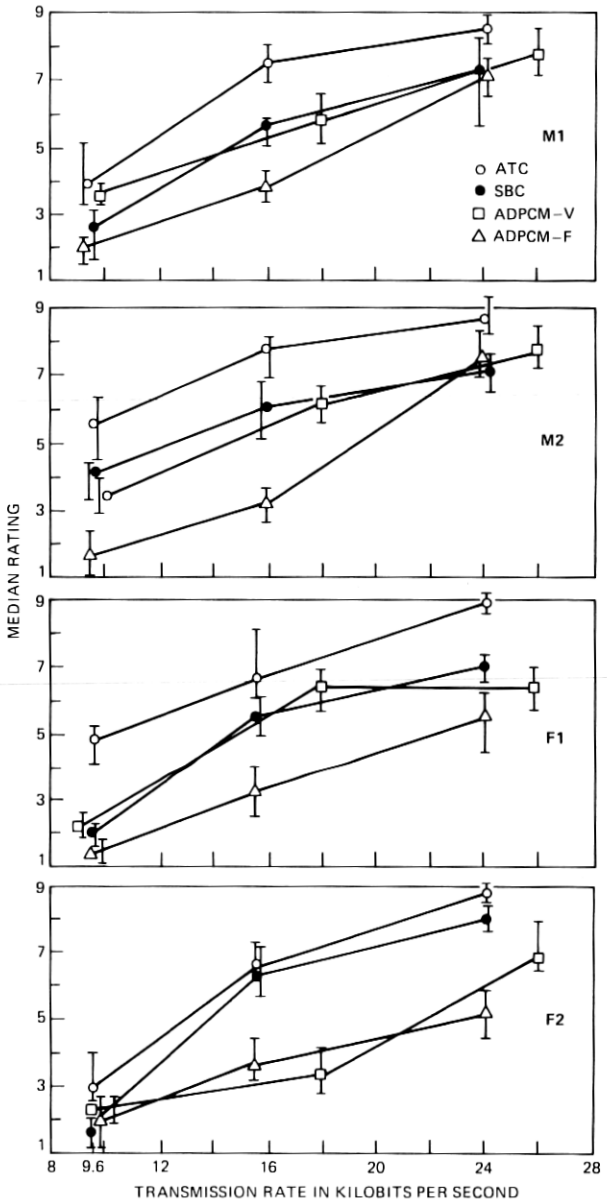


Fig. 6—Median rating of 12 coders for each talker.

ms) segments and has been found to be a more accurate predictor. In Fig. 8, the segmental s/n ratio, averaged across talkers, is plotted for each coder as a function of bit rate. These plots show that segmental s/n ratio is linearly related to the bit rate for each type of coder except

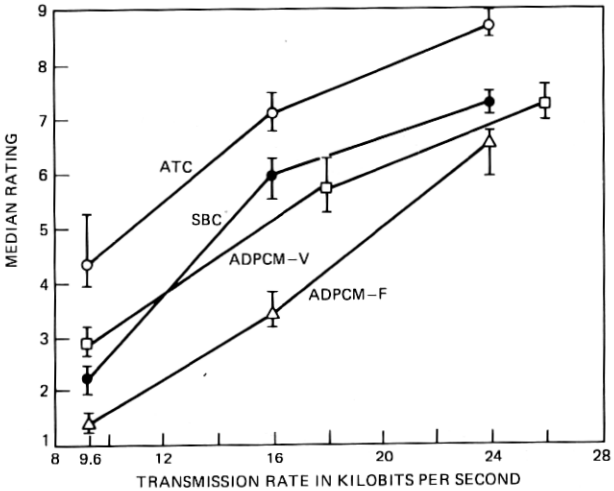


Fig. 7—Median rating of 12 coders (65 listeners × 4 talkers).

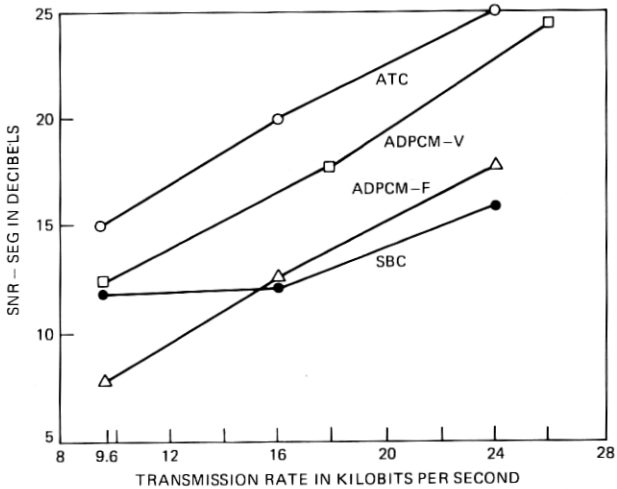


Fig. 8—Signal-to-noise ratio measured segmentally (15–20 ms) as a function of transmission rate for 4 types of coders.

SBC at low bit rates. While it orders ATC and the two ADPCM coders relative to each other at each bit rate, it underestimates the quality ratings of SBC shown in Fig. 7. In an effort to improve the effectiveness of s/n ratio as an indicator of quality, a number of frequency-weighted s/n ratio measures were computed, but they also were not highly successful in predicting the relative ordering of these four coders.[1,10]

4.2 Noise-to-signal measure

Another functional form for measuring the noise-to-signal power ratio in coders has been recently proposed[1] and analyzed.[10] This

measure was derived through concepts of log likelihood ratios and has the following functional form:

$$\bar{l}_m = 10 \log_{10} \left[1 + \sum_{j=1}^{B} e_j^2 / s_j^2 \right], \tag{2}$$

where the summation is taken over $B = 16$ frequency bands spaced according to the articulation bands in the range of 200–3200 Hz.[1,10] The bar above the equation denotes that $\bar{l}_m$ is the result of an average over (20 to 30 ms) segmental measurements, where e_j^2 is the segmental noise power and s_j^2 is segmental signal power in band j.

The values of $\bar{l}_m$, averaged over talkers, are plotted as a function of bit rate in Fig. 9a. This measure is a better predictor of the relative

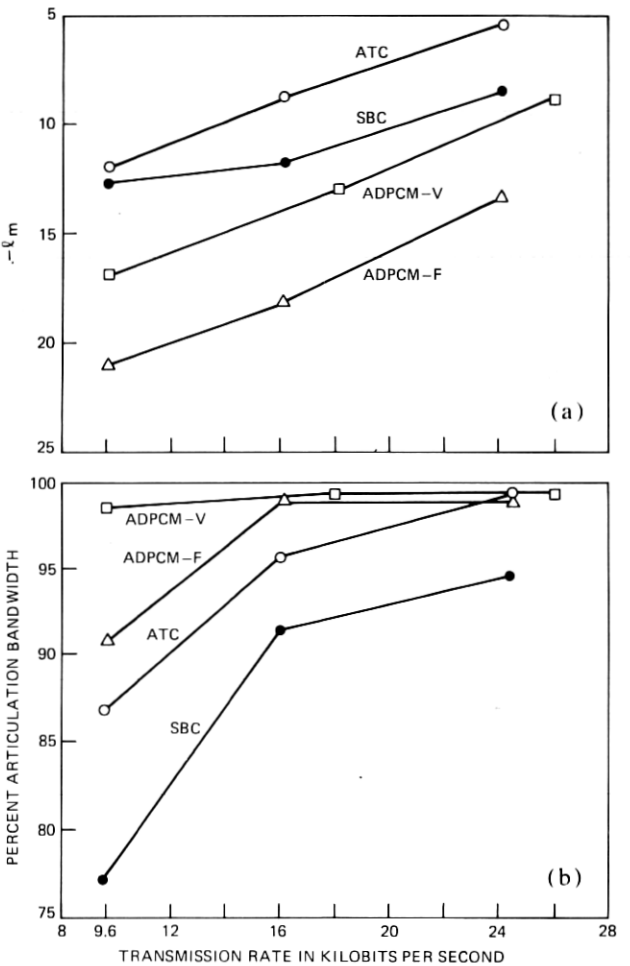


Fig. 9—(a) Noise-to-signal measure averaged over 16 articulation bands. (b) Percent of articulation bandwidth.

ordering of the coders at 24 and 16 kb/s where the distortions are primarily due to noise effects. At 9.6 kb/s, additional distortions due to bandwidth limitations are introduced by some of the coders. These distortions are not accounted for by the l_m measure and do not predict the subjective ratings at this bit rate.

4.3 Percent bandwidth

A bandwidth measure was defined[1,10] to account for the loss of bandwidth by the coders at low bit rates. The measure is defined as a simple percentage of the bandwidth on an articulation scale. Thus, the loss of a fixed bandwidth at a lower frequency has more importance than at a higher frequency. A coder with a flat frequency response from 200 to 3200 Hz is defined as one with a 100-percent articulatory bandwidth.

The bandwidth is measured on a segmental (32 ms) basis, and a smoothed spectral estimate of the input and output of the coder is computed for each segment. If, at any frequency point, the output power is less than, say, 10 dB of the input power, it is counted as a loss of bandwidth and is weighted according to the articulation scale. An average bandwidth is then computed over the entire utterance.

The percent bandwidth, plotted in Fig. 9b, complements the l_m measure in that it is sensitive to differences at 9.6 kb/s and relatively insensitive to differences at 16 and 24 kb/s. This measure correctly indicates that SBC at 9.6 kb/s has the lowest bandwidth due to the spectral gaps depicted in Fig. 2b. It also reflects the lower bandwidth of ATC at 9.6 kb/s due to the dynamic bit allocation in the algorithm and the 0 to 2800 Hz low-pass filter for ADPCM-F at this bit rate.

4.4 A combined measure

By forming a linear combination of the l_m and B_w (bandwidth) measures, a combined measure, Q, which accounts for both noise and bandwidth effects, can be defined. This measure has the form

$$Q = A_1 + A_2\, l_m + A_3\, B_w, \tag{3}$$

where A_1, A_2, and A_3 are computed by multiple regression techniques. The results of this combined measure for $A_1 = -3.78$, $A_2 = -0.42$, and $A_3 = 0.15$ are shown as the abscissa in Fig. 10, and the median ratings (of the 12 coder-bit rate combinations) as the ordinate. The solid dots in the scatterplot are the median ratings of the listeners summed over talkers and are the values used to compute the weighting coefficients. The four Δ's around each point are the median ratings for each of the four talkers. The high correlation (0.98) shows that this combined measure is a good predictor of the median ratings but will seriously err in predicting the ratings of some specific talker-coder combinations, particularly at low bit rates.

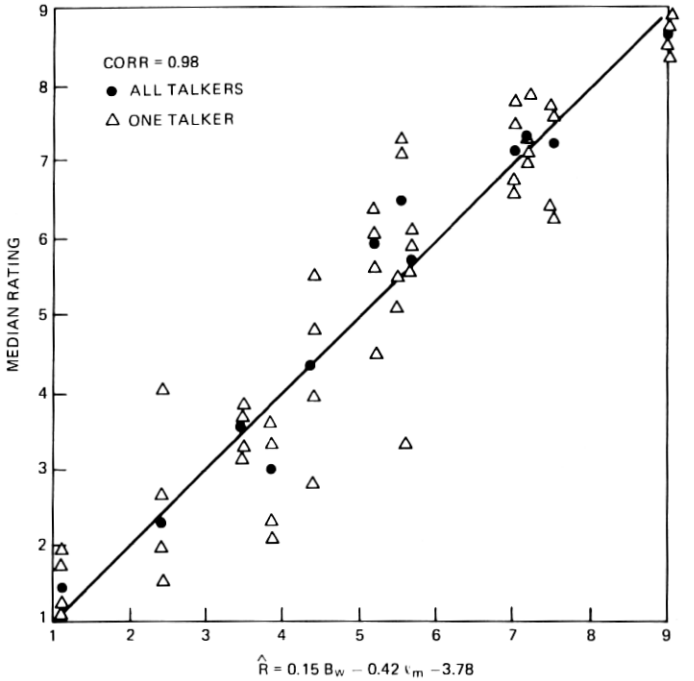


Fig. 10—Prediction of median rating by weighted combination of noise-to-signal and percent bandwidth measures.

V. DISCUSSION

Although the large variability in the ratings due to the different talkers and listeners rendered the quality ratings less precise than anticipated, an underlying pattern of relationships can be detected. ATC has the highest rating and ADPCM-F the lowest rating, regardless of bit rate, for most talkers and listeners. If cost/complexity is of no concern, then ATC is the most attractive of these coders. If cost/complexity is of concern, then SBC is an attractive choice since it is only slightly more complex than the ADPCM-F coder and about equal in quality to the costlier ADPCM-V coder.

The combination of two objective measures, l_m and B_w, each accounting for a different type of degradation, appears to be promising as a predictor of subjective ratings. Its precision is obscured by the interactive effects of the talkers and listeners. Of these two effects, the talker interaction is probably the easier to eliminate. The characteristics of different voices can be studied, and possibly the factors affecting the different coders could eventually be identified. However, the listener variability is a more difficult problem. Digital coding techniques are producing a variety of new and different types of degradations. As the bit rate is reduced, ATC produces a burble, SBC produces a reverberant quality, and the ADPCM coders produce a

signal-dependent noise. The variability in the subjects' ratings indicates that the trade-off for these different types of degradations is not the same for all people.

REFERENCES

1. R. E. Crochiere, L. R. Rabiner, N. S. Jayant and J. M. Tribolet, "A Study of Objective Measures for Speech Waveform Coders," Proceedings of the 1978 Zurich Seminar on Digital Communications, Zurich, Switzerland, March, 1978, pp. H1.1–H1.7.
2. P. Cummiskey, N. S. Jayant, J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52,* No. 7 (September 1973), pp. 1105–1118.
3. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM quantizers," Proc. IEEE, *62* (May 1974), pp. 611–632.
4. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Subbands," B.S.T.J., *55* No.10 (October 1976), pp. 1069–1085.
5. R. E. Crochiere, "On the Design of Sub-Band Coders for Low Bit Rate Speech Communication," B.S.T.J., *65,* No. 5 (May–June 1977), pp. 747–770.
6. P. Noll, "Non-adaptive and adaptive differential pulse code modulation of speech signals," Polytechnisch Tijdschrift, Den Haag, 1972, No. 19, pp. 623–629.
7. P. Noll, "Untersuchungen zur Sprachcodierung mit adaptiven Prädiktionsverfahren," NTZ, *27* (1974), pp. 67–72.
8. R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," IEEE Trans. Acoust. Speech and Sig. Proc., *ASSP-25* (August 1977), pp. 299–309.
9. P. W. Noll, "Adaptive Quantization in Speech Coding Systems," Int. Zurich Seminar on Digital Communication (October 1974), pp. B3.1 to B3.6.
10. J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A Study of Complexity and Quality of Speech Waveform Coders," Proc. 1978 IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., April 10–12, 1978, Tulsa, Okla. pp. 586–590.