Systems Division BELL-NORTHERN RESEARCH

TECHNICAL MEMORANDUM

TM 32055 October 1979

Comparative Evaluation of Residual-Excited Linear Prediction and Sub-Band Coding for Speech Transmission at 9.6 kb/s

79-14

PROPRIETARY

Copy #38

Author(s): D.C. Stevenson
P. Kabal

recommended by: Dale C Stevenson
Peter Kabal

approved by:

authorized by:

Dept. : 3R20

Keywords : Speech Coding

Abstract : Two coding schemes, Residual-Excited Linear Prediction (RELP) and Sub-Band coding were evaluated at a simulated transmission rate of 9.6 kb/s. The results of subjective testing indicates that the speech quality of these codecs is equivalent to approximately 4 bits log PCM. The sub-band codec offers a simpler coding scheme, whereas the RELP offers slightly higher quality. Both codecs produce highly intelligible speech.

Funder : Bell Canada
: B.P. Nicholls
OtherRef : Project B79-404

Case : 5514

bnr499a

SUMMARY

Two techniques for coding speech at 9.6 kb/s have been simulated and evaluated on a variety of sentences. The results indicate that both residual-excited linear prediction (RELP) and SUB-BAND coding provide for highly intelligible but somewhat distorted speech transmission. The best quality achieved with SUB-BAND coding is comparable to 4 bit PCM. Improvements to RELP coding indicate that somewhat higher quality, namely that equivalent to 4.5 bit PCM is attainable. This small quality difference is outweighed to a large extent by the additional computational complexity (and therefore cost) of RELP coding beyond that of SUB-BAND.

The quality evaluations carried out allow us to place RELP and SUB-BAND in the context of other speech coding techniques as shown in Fig. A. At bit rates below 4 kb/s conventional LPC coding offers the best quality. 9.6 kb/s transmission rate lies near the crossover between RELP and the SUB-BAND. At higher rates SUB-BAND is clearly preferable, while at lower rates RELP yields more favorable results. Above 16 kb/s toll quality can be attained with adaptive predictive coding or possibly with adaptive transform coding. The simulation and evaluation work reported

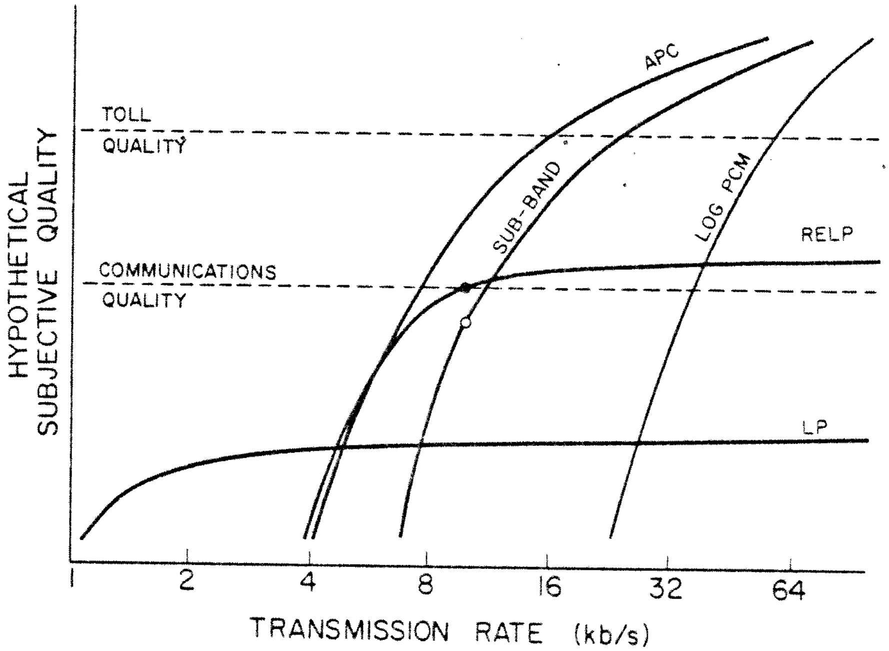

Fig. A Estimated subjective quality of several coders. APC = Adaptive Predictive Coding; LP = pitch excited Linear Prediction; RELP = Residual-Excited Linear Prediction. The data symbols (● = RELP, ○ = SUB-BAND) represent the results of the quality evaluations presented in this report.

here allows us to recommend the most appropriate coding technique for any application with specified transmission rate and quality constraints.

For speech transmission over data networks little significant quality improvement is attained by increasing the RELP bit-rate from 7.2 to 9.6 kb/s. However, where rates higher than 9.6 kb/s are feasible, significant quality increments result from the use of sub-band coding. Sub-band coding at 16 kb/s is currently under consideration for digital mobile radio applications. RELP or LP coding is recommended only for applications where the bit-rate economies are overriding and quality requirements can be relaxed, such as some specialized private networks.

Although a definite improvement over previous RELP simulations has been attained, the results suggest that attempts to further improve the classical RELP technique will not prove significantly productive. The limitations appear to be our inability to regenerate the high frequency speech components with sufficient accuracy. Future work, however, is suggested in exploring the optimal width of the residual band and how increased residual bandwidth can be traded off against more economical coding of the residual to achieve further quality improvements at the same bit rate.

Further explorations of sub-band coding are warranted, particularly in the area of improvements attained through time-varying assignments of bits to the individual channels as compared to fixed bit assignments. The robustness of RELP and sub-band to transmission errors must also be given further consideration.

A separate report, TM32057, addresses the quality APC coding near 16 kb/s. Quality only slightly inferior to toll quality has been achieved. The rate at which APC deteriorates as the bit-rate is reduced towards 9.6 kb/s is not well known today. A combination of sub-band and APC techniques may result in quality higher than that with any one technique alone. Improvements in this area will form the subject of continued work on this case.

DISTRIBUTION LIST

Bell Canada

1.	B.P. Nicholls	Dir. Dev. and Standards,	HQTD, 220 Laurier
2.	G.P. Strange	Asst. Dir. Dev. and Standards,	HQTD, 220 Laurier
3.	D.F. Barr	Dir. Technology Programs,	HQTD, 220 Laurier
4.	*J.A. Harvey	AVP Technology,	HQTD, 220 Laurier
5.	*J.R. Barry	Dir. Networks and Standards,	HQTD, 220 Laurier
6.	*J.W. Fraser	AVP Business Dev,	HQBD, F5, 25 Eddy, Hull
7.	L.G. Wilson	Dir. Business Dev,	HQBD, F5, 25 Eddy, Hull
8.	D.A. Carruthers	CCG, F9, 160 Elgin, Ottawa	

Bell-Northern Research

9.	*D.A. Chisholm	0008	Corkstown
10.	*J.S. Bomba	9B10	Central
11.	*R. Kennedi	3A00	Central
12.	*A.L. Brosseau	3B00	Montreal
13.	*R.B. Hosking	3C00	Toronto
14.	*S. Young	3D00	Central
15.	*A. Beauregard	3G00	Montreal
16.	*G.B. Thompson	3H00	Central
17.	*M. Kuhn	9B00	Central
18.	*J.B. Leger	3M00	Montreal
19.	M.L. Blostein	3R00	Montreal
20.	*A. Vennos	3S00	Central
21.	*P.G. Turner	3W00	Central
22.	*J.F. Tyson	3Z00	Central
23.	*J. Elliott	1A00	Corkstown
24.	*L.C. Beaumont	7A00	Meriline Court
25.	*G.C. Smyth	7D00	Meriline Court
26.	*B.R. Smith	7M00	Meriline Court
27.	C. Staples	6710	Palo Alto
28.	B. Prasada	3R10	Montreal
29.	S. Cohn-Sfetcu	3Z60	Central
30.	M. Ferguson	3R30	Montreal
31.	F. Daaboul	3R20	Montreal
32.	J. Turner	3R20	Montreal
33.	S. Hussain	3D45	Central
34.	M. Hunt	3R20	Montreal
35.	M. Lennig	3R20	Montreal
36.	V. Gupta	3R20	Montreal
37.	D. O'Shaughnessy	3Q00	Montreal
38.	P. Kabal	3Q00	Montreal

* Executive Summary

DISTRIBUTION LIST (CONT'D)

38.	P. Mermelstein	3R20	Montreal
39.	E. Dubois	3Q00	Montreal
40.	S. Sabri	3Q00	Montreal
41.	D.C. Stevenson	3R20	Montreal
42.	T.I.C. (2)	8E20	Central
43.	T.I.C. (2)	8E24	Montreal
44.	T.I.C. (1)	8E20	Toronto
45.	Spares (3)		

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

REFERENCES

1. Mermelstein, P., Kabal, P. and D. O'Shaughnessy, "Simulation of Digital Coding Techniques for Speech Transmission at 9.6 kb/s", TM 32031, Bell-Northern Research, Dec. 1978.

2. J.D. Johnston, "Digital Transmission of Commentary-Grade (7 kHz) Audio at 56 and 64 kb/s", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Wash. D.C.: 442-444 (1979).

3. W.I. Manson and D.W. Stubbings, "A Digital Split-Band Compandor for 64 kbit/s Coding of Speech with a 7 kHz Bandwidth", BBC Report RD 1977/41, Nov. 1977.

4. Un, C.K. and D.T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbits/s", IEEE Trans. Communications, COM-23, No. 12, pp. 1466-1474 (1975).

5. Markel, J.D., Gray, A.H., Jr., and H. Wakita, "Linear Prediction of Speech - Theory and Practice", Speech Comm. Res. lab., Santa Barbara, Calif. SCRL Monograph 10, (1973).

6. Atal, B.S., and S.L. Hanauer, "Speech Analysis by Linear prediction of the Speech Wave", Jour. Acoust. Soc. Amer. 50: 637-555 (1971).

7. Weinstein, C.J., "A Linear Prediction Vocoder with Source Excitation", Proc. EASCON '75: 30A-30G, Sept. 29 to Oct. 1, 1975, Washington, D.C., (1975).

8. Markel, J. and A. Gray, Linear Prediction of Speech, Springer-Verlag: Berlin, (1976).

9. Esteban, D.J., Galand, C., Maudit, D. and J. Menez, "A 4800 bps Voice-Excited Predictive Coder (VEPC) Based on Improved Baseband/Sub-band Filters", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Wash., D.C.: 975-979 (1979).

10. Viswanathan, R. and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems", IEEE Trans. Acoust., Speech and Sig. Proc., ASSP-23: 309-321 (1975).

11. Makhoul, J. and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Wash., D.C.: 428-431 (1979).

12. Dankberg, M.D. and D.Y. Wong, "Development of a 4.8-9.6 kbs RELP Vocoder", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Wash. D.C.: 554-557 (1979).

13. Viswanathan, R., Russel, W. and J. Makhoul, "Voice-Excited LPC Coders for 9.6 kbs Speech Transmission", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Wash. D.C.: 558-561 (1979).

14. Atal, B.S. and N. David, "On Synthesizing Natural-Sounding Speech by Linear Prediction", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Wash., D.C.: 44-47 (1979).

15. Kojima, H., Gould W.J. and A. Lambiase, "Computer Analysis of Hoarseness", Paper Presented at the Meeting of the Acoustical Society of America, Spring, 1979, Boston, Mass.

16. D. Esteban and C. Galand, "32 kbps CCITT Compatible Split Band Coding Scheme", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Tulsa, Oklahoma: 320-325, April 1978.

17. D. Sloan, "Adaptive Transform Coding of Speech", INRS-Telecommunications Technical Report 79-05, June 1979.

18. N.S. Jayant, "Adaptive Quantization with a One-Word Memory", BSTJ, Vol. 55, pp. 1119-1144, Sept. 1973.

19. R.E. Crochiere, "On the Design of Sub-Band Coders for Low-Bit-Rate Speech Communication", BSTJ, Vol. 56, May/June 1977, pp. 747-770.

20. R.E. Crochiere, "A Mid-Rise/Mid-Tread Quantizer Switch for Improved Idle-Channel Performance in Adaptive Coders", BSTJ, Vol. 57, pp. 2953-2955, Oct. 1978.

21. IEEE Recommended Practice for Speech Quality Measurements, IEEE Standard 297, April 1969.

22. Finney, J.D., Probit Analysis, Cambridge University Press, Cambridge, Mass. (1964).

23. Nakatsui, M., "A Unified Subjective Measure of Overall Speech Quality for Digital Waveform Coders", INRS Technical Memo (to appear).

24. Flanagan, J.L., Schroeder, M.R., Atal, B.S., Crochiere, R.E., Jayant, N.S. and J.M. Tribolet, Speech Coding", IEEE Trans. Comm. COM-27, No. 4: 710-736 (1979).

25. Reticon Corp., Data Sheet for the R5602 Transversal Filter.

26. TRW, Data Sheet for the TDC1010J Multiplier-Accumulator.

27. D. Esteban and C. Galand, "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes", IEEE Int. Conf. on Speech, Acoust. and Sig. Proc., Hartford, Conn., pp. 191-195, May 1977.

1. INTRODUCTION

Speech transmission at 9.6 kb/s is a requirement for operating speech communication links over contemporary digital networks. This report concerns an evaluation of two of the available low bit rate coding algorithms, Residual-Excited Linear Prediction (RELP) and Sub-Band coding. Both codecs have been investigated in the literature, but little information is available concerning subjective estimates of the quality of speech they produce. This report summarizes attempts to optimize the quality of these coders by analyzing and optimizing the individual signal processing operations which contribute to the resultant speech quality. Improvements have been incorporated where possible.

Linear Prediction (LP) has been extensively investigated in recent years as one of the more promising candidates for low bit rate speech transmission. In linear predictive coding, a short segment of speech (typically 25 ms) is fitted to an all-pole model in order to extract the short-time spectral information contained in that speech segment. In pitch-excited linear prediction, a local estimate of the pitch period is made, and this pitch estimate is transmitted together with the spectral information. This process allows transmission at very low bit rates (1 to 2 kb/s), but with limited intelligibility and naturalness. (See Mermelstein et al. [1] for a recent evaluation). RELP, on the other hand, attempts to bolster the quality of speech by encoding and transmitting the residual, i.e., the difference between the original speech waveform and the linearly-predicted waveform. This improves the naturalness of the speech, but only at the expense of a much higher total transmission rate. Research in RELP coding schemes has primarily been concerned with ways to encode and decode the residual to achieve transmission rates of 9.6 and 4.8 kb/s.

SUB-BAND coders use separate waveform coders in each of several frequency bands. In so doing, the perceptual effect of quantization noise is mitigated because of the different tolerances to noise in different frequency ranges. The waveform coders for the perceptually more important lower frequencies are assigned more transmission capacity than the higher frequency regions. Thus the regions which define the harmonic structure of voiced speech are relatively better reproduced. By carefully choosing the frequency span of the sub-bands and the allocation of bits to the individual sub-bands, sub-band coders can be designed for rates from 7.2 kb/s through to 64 kb/s. The lower rate produces speech that is distorted, yet very intelligible, while the upper rate has been applied to very high quality Commentary-Grade Audio [2],[3].

2. CONCLUSIONS

The performance of RELP and SUB-BAND coders has been evaluated at a simulated transmission rate of 9.6 kbits/second. Fine-tuning of the parameters which control the performance of these codecs shows that speech coded at this transmission rate is quite intelligible and has potential applications in communication systems where low bit rates and high intelligibility are the prime requisites. Preference tests using naive English-speaking subjects show that the preference rating of SUB-BAND is equivalent in overall acceptability to approximately 4 bit log PCM speech. Despite the fact that the RELP and SUB-BAND speech are characterized by distortions which are perceptually distinct from that of log PCM encoding, these results seem to be reliable statistical estimates which can be extrapolated to the general English-speaking population. They also concur with the assessments of experienced listeners in informal listening tasks.

RELP speech is characterized by a roughness which, although possessing a definite noise-like quality, is nonetheless distinguishable from the whiter noise-like quality of log PCM speech. Results of the present research suggest that it is possible to trade off the roughness for a somewhat "metallic" quality which appears to be more acceptable. Linguistically naive subjects rate this version of RELP (referred to throughout this report as "RELP(2)") at approximately 4.5 log PCM equivalent. Evidently, this RELP speech is at or near communications quality. However, this version of RELP speech still suffers from significant spectral distortions which want correction. Nonetheless, it is of significant interest that this improvement in RELP speech has been achieved with an inconsequential increase in algorithm complexity.

The SUB-BAND speech is both band-limited and, in the 4-band coder assessed in this report, contains gaps in the speech spectrum at 1000 and 1700 Hz. While apparently not significantly interfering with intelligibility, the combination of the band-limiting 300-3200 Hz and the presence of spectral holes imparts a definite hollow quality to SUB-BAND speech. An additional and evidently more serious signal degradation is a noticeable "warbling" which results from the extreme quantization of the sub-bands needed to achieve 9.6 kb/s. The results of the subjective preference tests and the comments of listeners lead one to believe that it is primarily this warbling quality which limits the acceptability of the SUB-BAND speech.

Even though the RELP and SUB-BAND speech are characterized by different perceptual distortions, the average placement along the log PCM scale appears to be consistent across groups of subjects, and substantially independent of both talker, sex of the talker, and the particular utterance which has been coded. The largest variation is due to the listener himself; different listeners react differently to the RELP and

SUB-BAND signal distortions, some preferring the RELP and some the SUB-BAND. The individual preferences range from approximately 3 bit log PCM equivalent to 5 bit log PCM. However, the average of 4.0 bit log PCM obtained in the present study appears to be a fair reflection of the general social acceptability of RELP and SUB-BAND speech in a simulated telephone communications context.

The RELP coder is substantially more complex than the SUB-BAND coder since it requires a least squares fit of a 20 ms section of speech signal to an all-pole model. This suggests that, ceteris paribus,the SUB-BAND coder is the simpler method for achieving 4 bit log PCM quality. However, probably little additional improvement can be expected from the SUB-BAND family of coders without a substantial penalty in complexity.

3. RESIDUAL-EXCITED LINEAR PREDICTION OF SPEECH

3.1 The RELP Codec

A brief synopsis of the Residual-Excited Linear Prediction (RELP) coding scheme is given in this section. For further details, the reader is referred to references [1] and [4].

The analysis phase of the Residual-Excited Linear Prediction (RELP) codec is shown schematically in Fig. 3.1. The first operation to be carried out is that of pre-emphasis, in which the input speech signal is spectrally flattened. If y(n) represents the pre-emphasized speech signal, then

$$y(n) = s(n) - c_p s(n - 1) \qquad (3\text{-}1)$$

where s(n) is the input signal. A value of $c_p = 0.95$ boosts the spectrum of the speech signal by approximately 6 dB per octave. During each frame (typically a 25.6 millisecond of speech signal), the LP parameters are estimated using the equation

$$\tilde{s}(n) = \sum_{k=1}^{P} a_k s(n-k) \qquad (3\text{-}2)$$

The predictor coefficients, a_k, are estimated in each frame typically using either the autocorrelation method [5] or the covariance method [6] such that the error signal, or residual, is minimized in a least-square sense. The residual, r(n), is given by

$$r(n) = s(n) - \tilde{s}(n) \qquad (3\text{-}3)$$

where $\tilde{s}(n)$ is the linearly predicted waveform. The major bit saving in LP (Linear Predictive) coding comes from the fact that P << N, where N is the number of points in a frame. (N is typically 200 points or so, and P is generally around 10).

The residual is derived from the speech signal by inverse filtering the input speech signal, using a filter defined by the LP coefficients a_k. Equation 3-3 above can be written as

$$r(n) = \sum_{k=0}^{P} b_k s(n-k) \qquad (3\text{-}4)$$

where $b_o = 1$ and $b_k = -a_k$, k=1,2,...,P. The residual so extracted is a waveform whose shape is reminiscent of that of the original speech signal. It has large excursions at the beginning of each pitch period, and thus retains the voicing information at that point in the speech waveform. If the frequency of the LP filter were to fit the spectral envelope of the signal exactly, the residual would consist of a train of spikes spaced at intervals of $1/f_o$, where f_o is the fundamental frequency of the speech waveform. In practice it is observed to

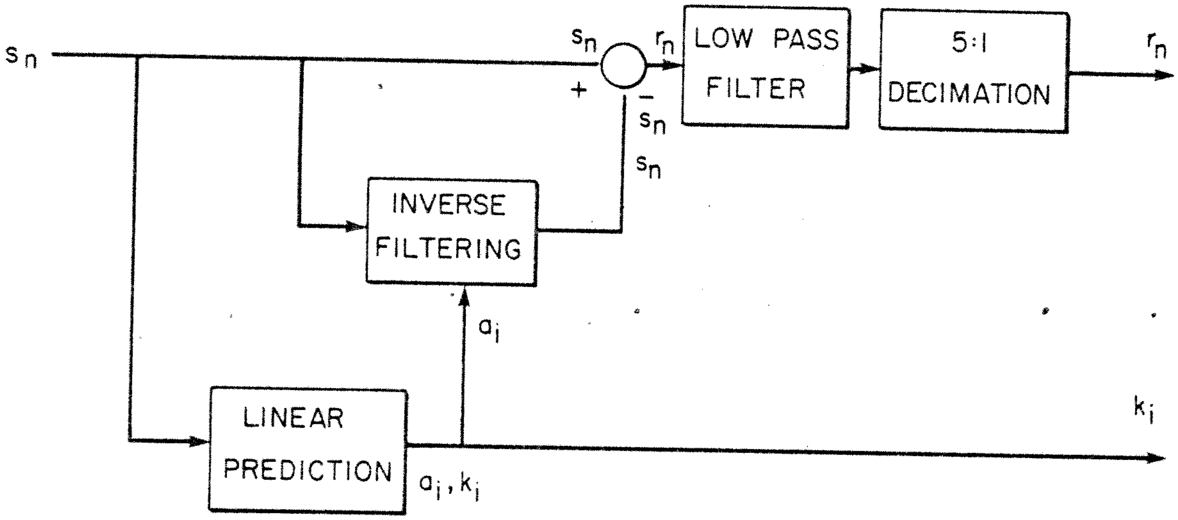

Fig. 3.1. Block diagram of RELP analyzer. During each frame, input data points s_n, n=1,...,N are analyzed, producing r_n, n=1,...N points of residual and k_i, i=1,...,P reflection coefficients. The a_k are the LP filter coefficients

retain a substantial amount of spectral information (so much so, in fact, that the original message is discernible in the residual waveform). RELP coding schemes attempt to optimize the quality of the synthesized speech signal by transmitting as much of this additional spectral information as possible. Indeed, if the full band r(n) is transmitted without quantization and is used as the excitation source for the synthesizer, the synthesized speech waveform is indistinguishable from that of the original.

To achieve low transmission rates with RELP coding, only a fraction of the information contained in the residual can be actually transmitted. It is difficult to take advantage of sequential redundancies in the residual since, by design, the residual is much whiter than the input speech signal. For instance, to achieve a transmission rate of 9.6 kb/s, at a sampling rate of 8000 Hz it is necessary to effectively encode the residual at approximately 1 bit per sample. The most commonly used method of achieving a substantial bit rate reduction is to transmit only the first few hundred Hertz (the baseband) of the residual. This low-pass filtering of the residual improves the sample-to-sample correlation, thus permitting adaptive coding schemes [4]. Alternatively, and evidently with equal efficiency [7], the residual can be decimated by a factor of L (i.e., only every Lth sample is transmitted), and each sample can then be coded using L bits. For a transmission rate of 9600 bits/second, a value of L=5 is used, in which case 8000 bits/second is used to transmit the residual waveform. This leaves 1600 bits/second available for transmitting the gain and spectral information.

Since only the baseband of the residual is transmitted, the first task during the synthesis phase is to reconstruct the high-frequency component (HFC) of the residual. This reconstructed residual, r(n), is then used to drive the LP synthesizer. In the Un and Magill paradigm [4], a full-wave rectifier is used to generate the higher harmonics from the baseband residual. Since this does not restore the higher harmonics to their proper spectral levels, this high frequency component is then spectrally flattened using a double-differencing operation:

$$y(n) = r(n) - 2G\ r(n - 1) + G^2 r(n - 2) \qquad (3\text{-}5)$$

The last stage in the reconstruction of the residual is to high-pass filter this newly-created HFC and add it to the baseband, adjusting the spectral levels of the baseband and the HFC as this is done. The resulting signal, r(n), has an average spectrum which is acceptably white, and for voiced segments demonstrates a harmonic structure which resembles that of the original residual. This residual is then used to drive the LPC synthesizer, and the result is an output signal whose spectral characteristics are similar to those of the input signal. A block diagram of the various operations in the synthesizer are shown in Fig. 3.2.

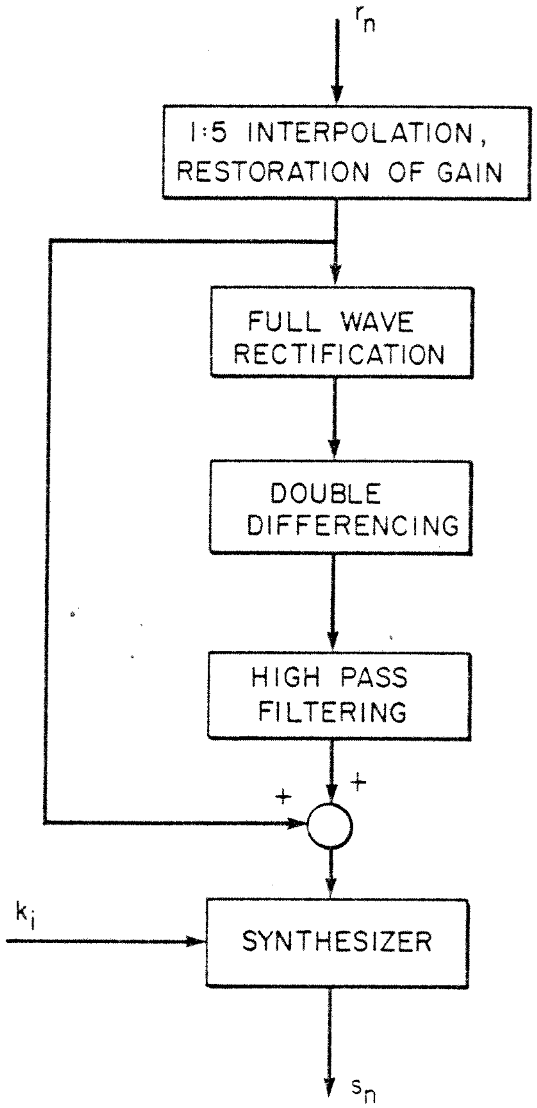

Fig. 3.2 Block diagram of RELP synthesizer. During each frame, N points of residual are received, r_n, n=1,...,N, as well as P reflection coefficients, k_i, i=1,...,P. N points of output signal, s_n, n=1,...,N are produced each frame

3.2 Factors Influencing RELP Quality

RELP-coded speech is generally characterized by the following distortions:

(1) frame-synchronous effects caused by quantization of the reflection coefficients, gain parameters and the residual

(2) frame-synchronous effects caused by the fact that the high-frequency component of the reconstructed residual only approximately duplicates that of the original residual

(3) a dominant strident quality

In the sections which follow, some of the implementation details of the software (FORTRAN IV) simulation of the RELP codec which affect the ultimate quality of the transmitted speech are described.

3.2.1 Generation of the Residual

The LP analysis shown in Fig. 3.1 was carried out by directly implementing the Markel and Gray autocorrelation subroutine [8, p. 219]. Eight poles (i.e., P=8) were used for the analysis. The analysis block was 25.6 milliseconds in length, and was advanced 20 milliseconds every frame, for a frame advancement rate of 50 times per second. This provided an overlap of 5.6 milliseconds for the LP analysis. The residual was produced by inverse filtering as described by Equation 3-3, except that the quantized versions of the a_k were used rather than the a_k themselves. This generation of the residual from the quantized rather than unquantized LPC coefficients is an important contributor to improving the quality of RELP speech, virtually eliminating the frame-synchronous effects caused by a mismatch between filter coefficients and the reconstructed residual [9].

3.2.2 Quantization and Transmission of the Residual

The result of the inverse filtering operation (Equation 3-4) is an array of M points, r(n), n=1,...,M, where $M=f_s/f_r$. f_s is the sampling frequency and f_r is the frame rate. In the present study, f_s was fixed at 8000 Hz and f_r was fixed at 50 frames/second. Thus, M=160, which corresponds to 20 milliseconds of speech. The analysis frame is N=204 points, or 25.6 milliseconds of speech. This provides 44 points of frame to frame overlap. Next, the residual was low-pass filtered using a 31-tap FIR filter with a cut-off frequency of approximately 800 Hz. At this time, two gain factors were recovered,

$$\sigma_L = \left(\frac{1}{N} \sum_{n=1}^{P} r_L(n)^2 \right)^{\frac{1}{2}} \tag{3-6}$$

and

$$\sigma_H = \left(\frac{1}{N} \sum_{n=1}^{P} r_H(n)^2 \right)^{\frac{1}{2}} \tag{3-7}$$

which represent the RMS values of the baseband residual ($r_L(n)$) and the high-frequency component ($r_H(n)$). (In the Un and Magill paradigm, only a single gain parameter is transmitted, and the baseband and HFC are combined in a fixed ratio). The baseband, $r_L(n)$ was set to unit variance in preparation for log PCM encoding:

$$z(n) = \frac{r_L(n) - \bar{r}_L}{\sigma_L} \tag{3-8}$$

The parameters $\sigma_L$and σ'_H, where

$$\sigma'_H = \sigma_H/\sigma_L \tag{3-9}$$

were transmitted for each frame. The offset,$\bar{r}_L$, was small and was not transmitted since informal listening tests showed that not restoring this offset did not introduce any significant additional distortions in the output speech.

Prior to transmission, the baseband was decimated by selecting only every 5th point. The effective number of points in the residual for one frame was thus reduced from 160 to 32. Each of these 32 residual points was then encoded by 5-bit log PCM for a total bit rate of 8000 bits/second. The fact that the residual had a zero mean on a frame-by-frame basis ensure that the residual signal was always properly positioned with respect to the quantizer levels. This effectively made the quantization of the residual,independent of the absolute signal level.

3.2.3 Quantization and Transmission of Coefficients

The coefficients transmitted for each frame are shown below in Table 3-1. The log area ratios, k_i,

$$k_i = \log \frac{1 + k_i}{1 - k_i} \tag{3-10}$$

were transmitted rather than the reflection coefficients, k_i. (The rationale for quantizing and transmitting the log area ratios rather than the reflection coefficients is provided in [10]. In essence, errors in transmission of the reflection coefficients cause spectrally more significant effects when the k_i close to either + 1 or -1. The log area ratios, on the other hand, tend to have more uniform spectral sensitivity, and hence are more suitable for transmission). A greater number of quantizer levels were used for the lower order coefficients since they carry the bulk of the spectral information. (The k_i form an ordered set of parameters, such that if $i \quad j$, then k_i is more important than k_j. This ordering of the k_i is an important property for efficient transmission of the spectral information [10]). Only 8 poles were used in the LP analysis in order to ensure enough bits were available for adequate quantization of the gain parameters.

Investigations in which 12 poles were used instead of 8 showed no significant improvement in speech quality.

The parameters "transmitted" for each frame and the number of levels assigned to each are shown in Table 3-1 below.

TABLE 3-1

Quantizer Levels Used in RELP Coding

PARAMETER	LEVELS	BITS
r_L	32	5
σ_L	64	6
σ'_H	16	4
k_0	32	5
k_1	16	4
k_2	16	4
k_3	8	3
k_4	4	2
k_5	4	2
k_6	2	1
k_7	2	1

The gain parameters and the residual (after reduction to unit variance via Equation 3-8) were quantized logarithmically with Mu=255. Since the residual was decimated by a factor of 5, 5 bits were available for encoding each sample indicated by the 32 levels shown in Table 3-1.

The gain parameter σ_L controls the absolute level of the output speech. Informal testing showed that it was mandatory to provide reasonably fine quantization of σ_L: fewer than 32 levels for the gain introduced perceptible noisiness into the output speech, as well as frame-synchronous chirps caused by sudden changes in the absolute level of the residual. In the present circumstances, 64 levels were provided to ensure adequate quantization. These levels were provided at the expense of finer quantization of the higher k_i which, as was pointed out above, carry little spectral load.

The second gain parameter, σ'_H , is a relative gain parameter (Equation 3-9) used in the synthesizer to restore the relative level of the HFC. Only 16 levels were provided for this parameter since its range of variation is not as large as that of σ_L. The average value of σ'_H was determined to be 2.0 on the basis of several test sentences but it varies an order of magnitude between frames. This parameter was transmitted to allow the spectral balance of the baseband and HFC to be restored for each frame independently, since deriving the residual from the quantized a_k tends to destroy the whiteness of the residual.

All transmitted parameters were mapped onto the quantizer levels by the linear transformation

$$y = ax + b \tag{3-11}$$

where x represents the value of one of the parameters listed in Table 3-1. To improve the performance of the quantizers, the scaling parameters a and b were optimized for each parameter by minimizing the sum of squares difference between the unquantized and quantized parameter values over a whole sentence. This optimization procedure was carried out on a speech data base consisting of four utterances spoken by each of five male and four female talkers. The mean of these scaling parameters a and b were then taken as being representative of the entire data set.

3.2.4 Reconstruction of the Residual

The stages in the reconstruction of the residual at the synthesizer are shown in Fig. 3.2. The residual was first upsampled 1:5 by inserting 4 zeroes between each of the transmitted points. Each of these points was then multiplied by $5\sigma_L$ to restore the residual to its pre-encoded value. The baseband was then recovered by filtering with an 800 Hz low-pass filter. Restoring the signal level prior to low-pass filtering minimizes the effects of quantization of the gain as well as lessening the effects of not restoring the offset $\bar{r}_L$(i.e., Equation 3-8).

A copy of this baseband was then used to recreate the high-frequency component. The first step was full-wave rectification, which regenerates the higher harmonics of the residual. (Various degrees of rectification in between full and half-wave rectification were tried but no difference was discernible in the output speech (see [5] also). Therefore, a fullwave rectifier was used for convenience). This rectification was followed by a double-differencing operation (Equation 3-5 with Ç=0.8) to spectrally flatten the signal. This signal was then high-pass filtered by a filter which was the mirror image of the low-pass filter used in the coder to extract the baseband. This regenerated high-frequency component was then restored to its original level by:

$$r_H(n) = \sigma_L \sigma_H' \, r_H(n) \Bigg/ \left(\frac{1}{N} \sum_{n=1}^{P} r_H(n)^2 \right)^{\frac{1}{2}} \tag{3-12}$$

The full-band residual was then obtained by adding the two together:

$$r(n) = (1-R)r_L(n) + Rr_H(n) \tag{3-13}$$

Since the HFC generation process described above restores the RMS value of the high frequency component, at this point both the baseband and the HFC have been restored to their absolute levels, hence, the theoretical value of R should be 0.5. To assess the possibility that it may be preferable to lower the spectral level of the HFC in order to reduce the perceptual salience of the roughness, an informal test was carried out in which R was varied between 0.2 and 5.0. Judgements by two trained listeners to randomly presented utterances produced a value of approximately R=0.5 as the optimum value. R was therefore set at 0.5.

3.2.5 Delaying the Residual

The residual produced from Equation 3-13 matches the original residual in approximate spectral balance and total RMS level. However, because of the various filtering operations carried out in the coder and in the decoder, the reconstructed residual is delayed with respect to the original residual. Consequently, the first N_d points of the residual of a given frame, as reconstructed by Equation 3-13, actually correspond to the LP coefficients for the previous frame. Therefore, in the synthesizer, insertion of the new reflection coefficients must be delayed until the N_dth point in the new residual is reached. For the 31-tap filters used in the present study, the cumulative delay imparted by the various filtering stages in the coder and decoder was N_d=47. Informal listening tests showed that values of N_d differing from this optimum value by more than 10 or so points produced detectable additional roughness in the output speech.

3.2.6 Interpolation of Reflection Coefficients

The coefficients of the lattice synthesizer are updated once per frame (as described above in Section 3.2.4), and at this point small transient effects are produced. These transients are heard as small "chirps" whose magnitude increases with the frame-to-frame differences in either the k_i or the gain and are especially noticeable if the input speech signal is restricted to a narrow spectral band. They can be eliminated by interpolating the k_i and gain parameters across the frame boundary. Informal listening tests showed that interpolation of the reflection coefficients is unnecessary if (a) the residual is derived using the quantized a_k and (b) the gain parameter, σ_L, is not quantized too coarsely. Under these conditions, for most talkers these frame-synchronous effects are no longer audible. This is no longer true at 4.8 kb/s where both σ_L and the k_i must necessarily be quantized coarsely.

3.2.7 Harmonic Structure of the Residual

The processing details described above in Sections 3.2.1 though 3.2.6 resulted in speech which showed virtually no effects of quantization of either the gain or reflection coefficients. Also, transforming the residual to unit variance prior to quantization minimized the effects of the quantization. Only a small additional amount of roughness could be attributed to the quantization of the residual to 5 bits. Thus, for a fixed bandwidth for transmission of the residual, the limiting factor governing the quality of the RELP speech is the roughness generated by the process used to regenerate the higher harmonics of the residual.

The origin of the roughness of RELP-coded speech is still unclear. Undoubtedly it arises from several sources, since various claims have been made in the literature concerning ways to reduce it. Dankberg and Wong, for instance, suggest that the roughness can be reduced by a suitable choice of filters for the low-pass and high-pass operations [12].

Viswanathan et al. [13], on the other hand, suggest that aliasing effects due to the non-linear distortion procedure would produce components between harmonics and that this is a source of roughness in the output speech. They propose upsampling the residual to twice the sampling fequency prior to the non-linear distortion process (i.e., the full-wave rectification). Presumably these aliasing components can then be eliminated by digital filtering, and the residual then downsampled to its original rate. This upsampling to $2f_s$ was implemented in the present study but informal listening tests by several observers verified that no significant changes occurred, a result in agreement with the assessment of Dankberg and Wong [12].

To test the possibility that the roughness arose from the overlapping of spectral components in the baseband and HFC, various high pass filters were tried (re Fig. 3.2). The "normal" algorithm requires a high pass filter with a cut-off frequency of 800 Hz to complement the filter used in the creation of the baseband. Several high-pass filters with cut-off frequencies in the range 800 to 1200 Hz were tried, and informal listening showed that this reduced the roughness of the output speech. Evidently, some, but not all, of the roughness is associated with overlapping of spectral components in the 800-1200 Hz region of the spectrum. However, the speech also acquired a slight hollow property, and thus the reduction in roughness was probably due primarily to the fact that part of the spectrum had been removed. As a further test, the low pass and high pass filters were changed to have steeper skirts, with a stop-band of approximately -40 dB. The roughness of the output speech was not observed to decrease, again suggesting that the roughness of the RELP speech is not due to overlapping of components across the 800 Hz filter boundary. The 800 Hz filter used for the RELP algorithm appears to be adequate.

The work of Atal and David [14] concerning restoration of the proper amplitudes and phases of the harmonics of the residual suggest that improving the short-term spectral shape of the residual also improves the quality of the output speech, although in their case at the expense of considerably more (pitch-synchronous) processing. Kojima et al. [15] report that horseness in natural speech manifests itself as an increased noise level between harmonics. Spectrograms of voiced RELP-coded speech produced by the Un and Magill paradigm show just such a clouded harmonic structure. Thus, it is possible that poor definition of harmonic structure is primarily responsible for the hoarseness of RELP speech.

Fig. 3.3 shows a comparison of (a) the high frequency component of the original residual (i.e., above 800 Hz), and (b) the high frequency component as reconstructed from fullwave rectification and spectral flattening. The difference observed in Fig. 3.3 is typical for voiced segments. The large spikes in the HFC evidently act as secondary sources of excitation in the synthesizer, and it is possible that in so doing they may introduce harmonic components with different phases. To modify the temporal structure of the HFC, the baseband was altered by a differencing operation prior to the rectification stage, viz,

$$y(n) = r(n) + ar_L(n-1) + br_L(n-2) \qquad (3-14)$$

The effect on the HFC for a=-1.4 and b=0.4 is shown in Fig. 3.3(c). (These values of "a" and "b" were determined by informal listening tests in which "a" and "b" were varied. As "b" goes to zero and "a" goes to -1, the operation becomes that of a simple differencing. As both "a" and "b" go to zero, the operation reduces to just y(n)=r(n). The acronym RELP(1) will be used to designate the RELP codec when a=b=0 and RELP(2) will refer to it when a=-1.4, b=0.4).

The following changes are observed in the spectra of the residual: first, a hole is placed in the spectrum around 900 Hz, attesting to the fact that the principal effect of the differencing operation (Equation 3-14 above) is a high pass filtering of the baseband. The average spectrum of an utterance coded by the RELP(1) and RELP(2) coders is shown in Fig. 3.4(a) and (b) respectively. The hole around 1200 Hz is clearly visible in Fig. 3.4(b).

The second effect of this additional differencing stage is a marked reduction in the roughness of the output speech. Spectral sections show that the harmonic structure of this version of RELP speech (henceforth referred to as RELP(2)) is generally much improved in the region 2000-4000 Hz, but is made somewhat worse for lower frequencies. The improvement in the output speech probably can be attributed to the "cleaning up" of the harmonic structure [15].

The third effect of this additional differencing stage is an overall "metallic" quality which is pronounced for some talkers and almost non-existent for others. Accompanying this metallic ring is a slight hollowness which probably arises from the hole in the spectrum around 1200 Hz.

Informal listening tests by several trained listeners indicated that in spite of the slight metallic quality of the RELP(2) stimuli, there was a noticeable improvement in overall quality compared to the RELP(1) stimuli

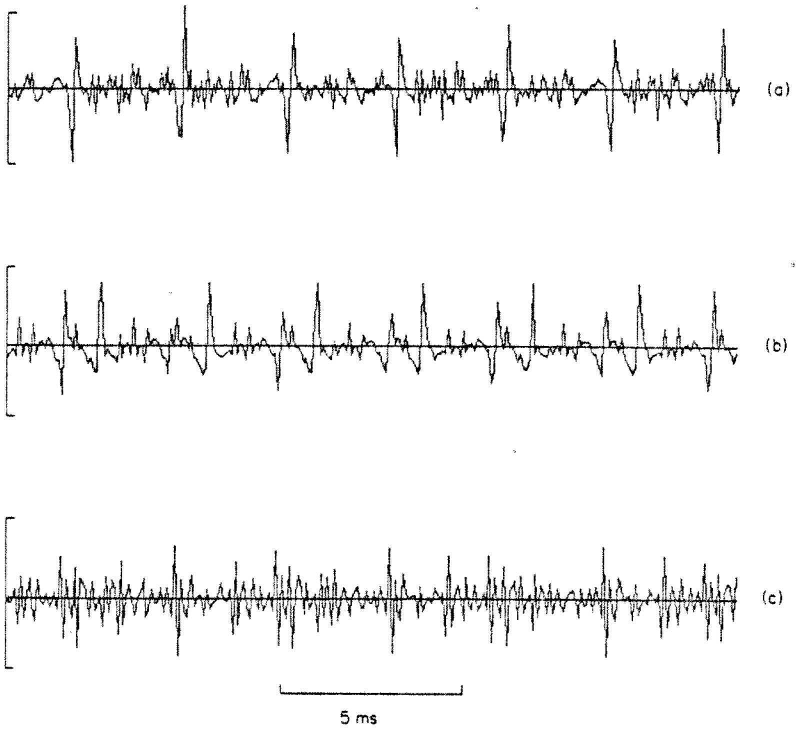

Fig. 3.3 Waveforms of high frequency component for one frame of voiced speech. (a) original (b) RELP(1) (c) RELP(2)

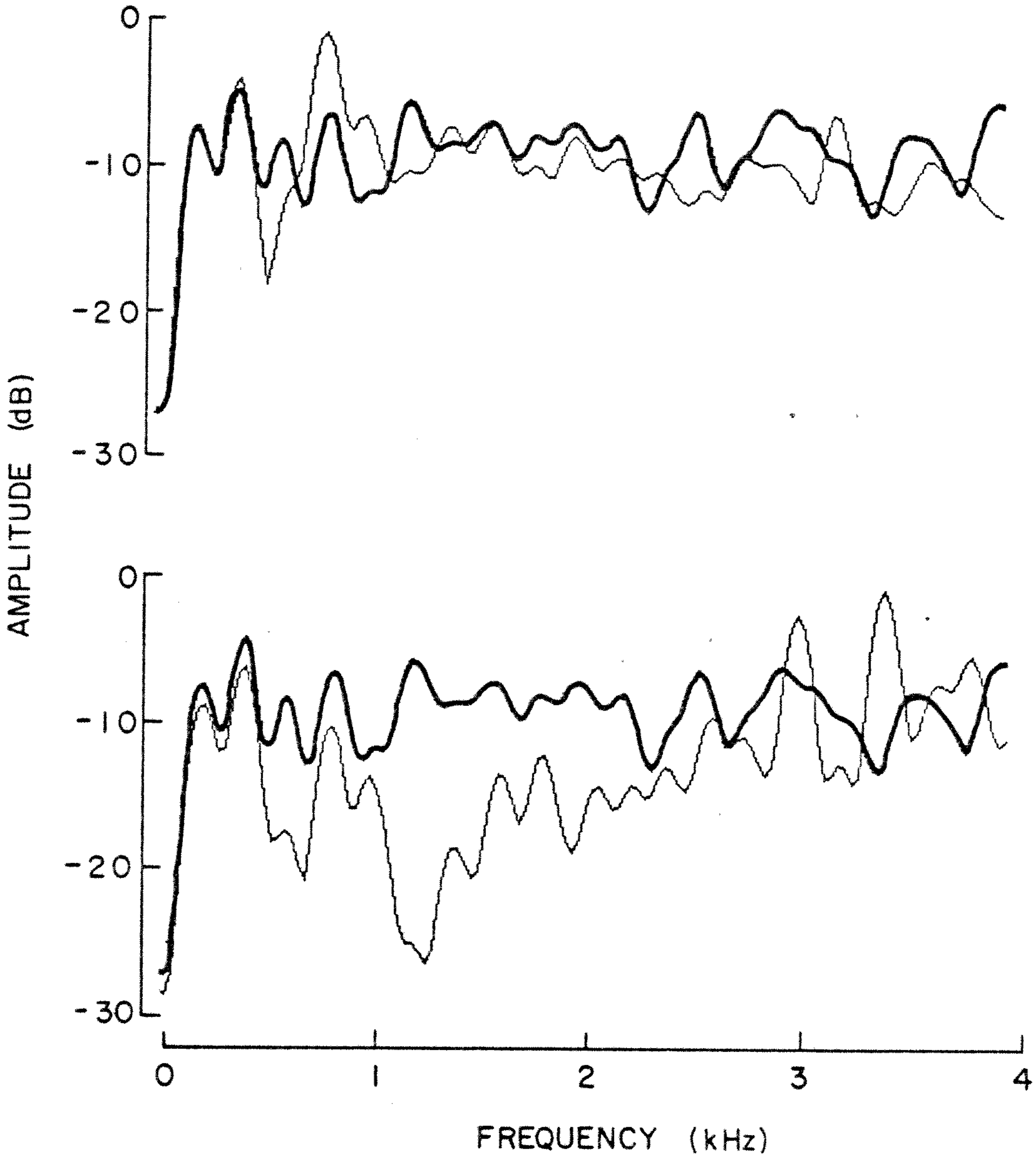

Fig. 3.4 Comparison of spectrum of original residual (heavy line) with coded spectra. (a) Original vs. RELP(1) (b) Original vs. RELP(2)

(i.e., stimuli prepared without this extra differencing stage). In the subjective testing described later (Section 5), both the RELP(1) and RELP(2) stimuli are evaluated, along with the SUB-BAND stimuli described below in Section 4. At present, the results for the RELP(2) stimuli are offered only as an indication of potential quality for RELP coding schemes; more work is necessary to (a) understand why an improved harmonic structure results, and (b) to restore the whiteness of the full band residual.

4. SUB-BAND CODING

In a SUB-BAND coder, the input speech is filtered into several non-overlapping spectral regions. The output of each sub-band filter is frequency translated to baseband and sampled. The samples of each sub-band are coded independently using an adaptive quantizer. The reconstruction process then follows by frequency translating each sub-band to its original position. These operations are shown in a block diagram in Fig. 4.1.

The following observations motivate the SUB-BAND technique.

(1) The independent quantization of the sub-bands results in quantization noise which is localized in frequency. Adjacent sub-bands have uncorrelated noise components. The overall effect is to greatly lessen the subjective impact of the quantization process.

(2) The width of each sub-band and the precision to which the signal in that sub-band is quantized can be adjusted to utilize the available transmission channel efficiently. High energy segments, generally the lower sub-bands, can be allocated more transmission capacity.

(3) The quantizers in each sub-band can individually adapt to the signal energy contained in the corresponding frequency band.

4.1 Factors Affecting SUB-BAND Quality

The major factors controlling the quality of SUB-BAND coded speech are the choice of the sub-bands and the coding of each sub-band. These factors are described separately below.

4.1.1 Number of Sub-bands

Assuming ideal filtering, an increase in the number of sub-bands will tend to improve the quality of the coded speech, though with diminishing returns after the first few sub-bands [16]. Practical considerations ultimately limit the number of sub-bands that can be used for the following reasons.

(1) The complexity of the coder increases as the number of sub-bands is increased. The major factor is the filtering needed for sub-band separation and reconstruction. This effect is exacerbated by the need to sharpen the filter skirts in the regions of overlap as the number of sub-bands increases.

(2) As the number of sub-bands increases, the sampling rate for each sub-band decreases. (The sampling rate for each sub-band is twice its bandwidth). The time evolution of the sample values in a sub-band can then no longer be used to get a good estimate of the short-time energy in that sub-band. The bandwidth (and hence

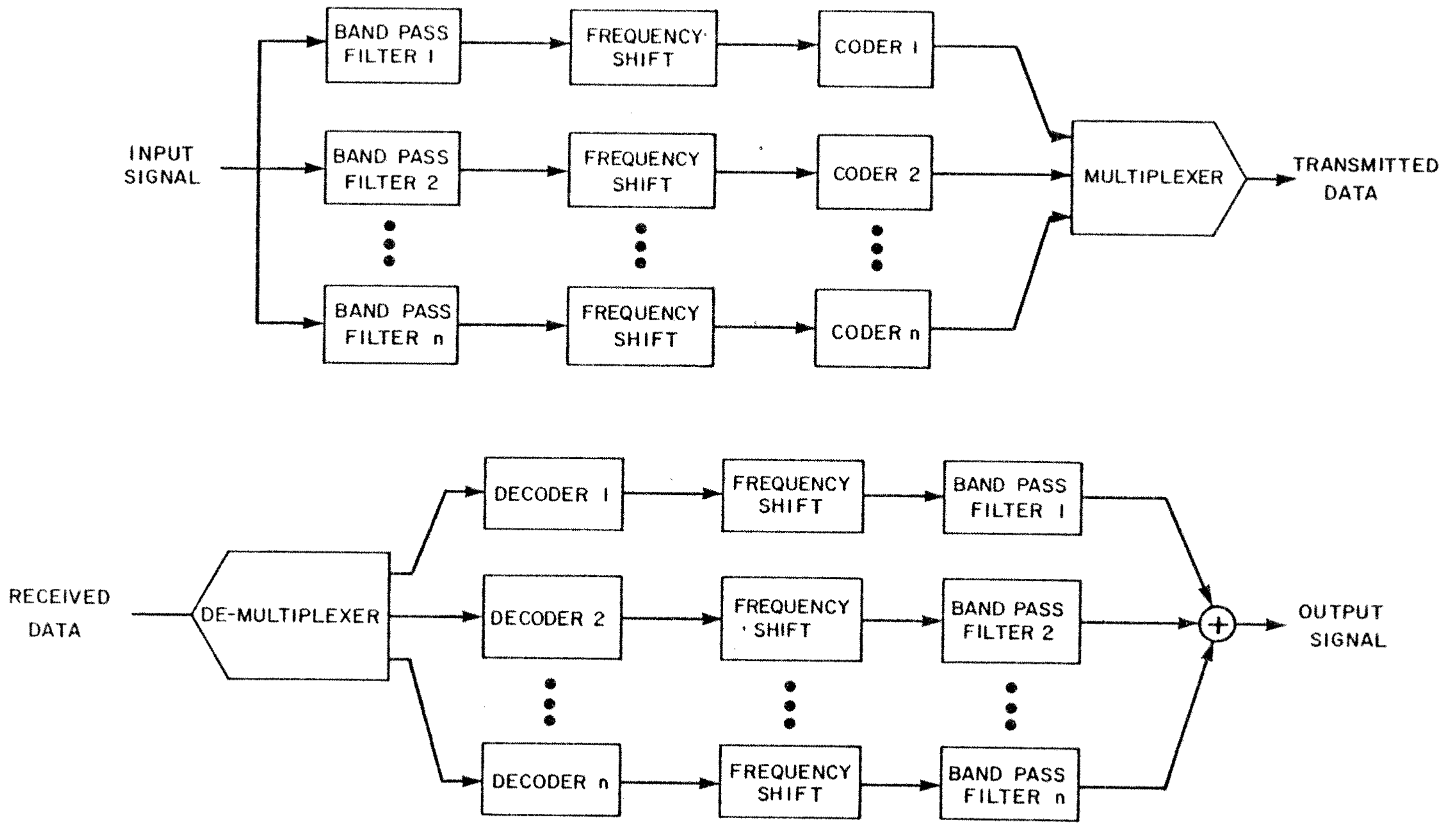

Fig. 4.1 A General Block Diagram for a Sub-Band Coder

sampling rate) of each sub-band should be large enough to allow the quantizer adaptation strategy (described below) to follow the signal energy trends in each sub-band at the syllabic rate.

(3) More critical filtering implies filters which involve a longer time delay. While moderate time delays do not effect the link performance per se, they do affect the performance of a duplex system with an echo path. Since sensitivity to echo increases with coding delay, coders used on the telephone network must minimize the coding delay incurred.

These considerations limit the number of sub-bands to between 4 and 8. Note that alternate techniques exist that allow the use of a large number of sub-bands using transform techniques [17]. The quantization strategy for these schemes is significantly more complex.

4.1.2 Sub-Band Quantizers

The coders for each sub-band use an adaptive quantizer proposed by Jayant [18]. This algorithm varies the quantizer step size as

$$\text{delta}(n+1) = \text{delta}(n)*\text{mult}(k) \qquad (4-1)$$

where delta(n) is the step size at time n, and mult(k) is the multiplier associated with the k'th output level. The multipliers are chosen such that if an inner level of the quantizer is used, the multiplier is less than unity, while if an outer quantizer level is used, the multiplier is larger than unity. 'With this choice of multipliers the range of the quantizer is expanded or contracted to match the amplitude range of the signal as the outer or inner levels, respectively, of the quantizer are exercised. The dynamic range of delta(n) is limited to 100:1. This prevents the step size from getting so small during a period of zero input that the recovery time becomes appreciable.

At the transmission rate chosen for this study, 9.6 kb/s, at least some of the sub-bands must be coded with near 1 bit per sample. Jayant's scheme no longer works with only two output levels. Goodman has proposed a 1+1/K bit quantizer (as reported in [19]). In this modified version of the Jayant scheme, every K'th sample is transmitted as two bits, one for sign and one to indicate amplitude. The samples in between are coded using one bit per sample, indicating sign only. (An alternate 1 bit per sample quantizer using adaptive delta modulation was slightly less effective).

An additional modification to the quantizer has been suggested by Crochiere [20]. As applied to this study, if the quantizer step size is within a factor of 2 of the minimum step size, a new quantizer is switched in. This new quantizer has an odd number of levels, and hence has as an output level the zero value. This modification reduces the idling noise and the low level tones in each sub-band. However this idling noise has a masking effect on other distortions. Without it, the other distortions now stand out more plainly. To help combat this, low level bandlimited white noise can be added to the coded signal. With the noise at a level 20 dB below that of the speech, some masking of the coding artifacts is achieved. (This masking noise was not employed in the subjective tests).

4.1.3 Sub-Band Filters

The filters used for separating the Sub-Band frequency ranges are linear phase finite impulse response (FIR) filters. With 255 taps each, these filters do not themselves introduce any perceptible distortion. Less complex filters with fewer taps or even perhaps infinite impulse response (IIR) filters (with their inherent phase distortion) would certainly suffice if implementational complexity is important.

4.1.4 Frequency Translation

An earlier report [1] outlined the alternatives for frequency translation. The scheme adopted uses a subsampling procedure to sample the sub-bands and simultaneously achieve a shift to baseband. While this technique entails some compromises as to the location and width of the sub-bands, the advantages of its simplicity far outweigh the restrictions imposed.

4.1.5 Configurations for Sub-Band Coders

4.1.5.1 Configuration 1

The first scheme tried used 5 sub-bands as shown in Fig. 4.2 and Table 4-1. The range of speech frequencies coded was restricted to between 200 and 3200 Hz. In keeping with the philosophy of SUB-BAND coding, the lower sub-bands are narrower and quantized with more precision than the upper sub-bands. However the overall effect of the very coarse quantization over a large band of frequencies is very noticeable. This configuration employing different bit allocations, is very effective for coding at higher rates, specifically at 16 kb/s.

4.1.5.2 Configuration 2

Crochiere [19] suggests that leaving gaps between the sub-bands is an effective way to allow more precision to be allocated to the sub-bands that are transmitted. Fig. 4.3 and Table 4-2 show the 4 sub-bands used. In the absence of quantization, the gaps give the speech a slightly hollow sound. With quantization, the better allocation of bits (compared to Configuration 1) reduces perceptual effect of the quantization noise. Overall, the trade-off of hollowness for less noise is effective. This

TABLE 4-1 QUANTIZER BIT ALLOCATION FOR FIVE-BAND CODER

SUB-BAND	FREQUENCY RANGE Hz	SAMPLING RATE Hz	BITS	RELATIVE STEP SIZE
1	160-320	320	3	0.6
2	267-533	533	3	1.0
3	480-960	960	2	0.25
4	960-1920	1920	1.33	0.07
5	1920-2880	1920	1.25	0.02

TOTAL RATE = 9439 b/s

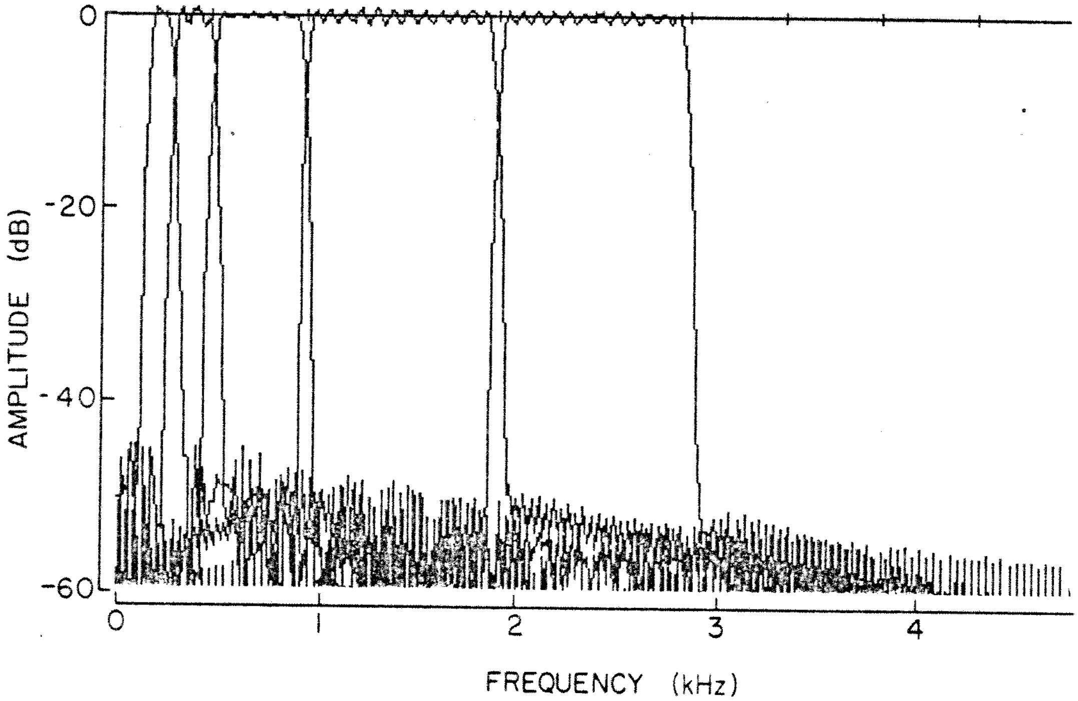

Fig. 4.2 Five-band frequency response

TABLE 4-2 QUANTIZER BIT ALLOCATION FOR FOUR-BAND CODER

SUB-BAND	FREQUENCY RANGE Hz	SAMPLING RATE Hz	BITS	RELATIVE STEP SIZE
1	240-480	480	3	0.6
2	480-960	960	3	1.0
3	1067-1600	1067	2	0.25
4	1920-2880	1920	1.5	0.07

TOTAL RATE = 9334 b/s

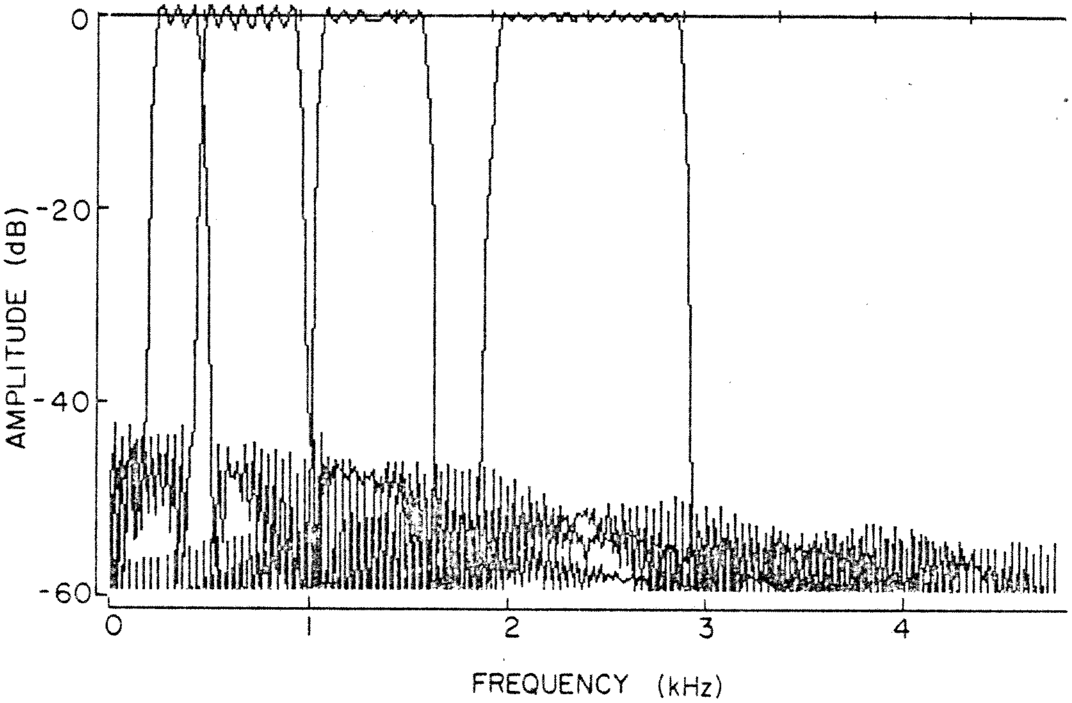

Fig. 4.3 Four-band frequency response

configuration, with suitable bit assignments [19], can be used at rates down to 7.2 kb/s. At this lower rate, the quality degrades, but the speech is still very intelligible.

4.1.5.3 Configuration 3

Subjective evaluations were carried out using Configuration 2 at 9.6 kb/s. During the course of this investigation, it was discovered that small amounts of aliasing confined to the band edges have little effect on the perceived quality. In this configuration, aliasing was used to fill in the gaps between the bands in Configuration 2. This was achieved by increasing the filter widths until the gaps disappeared. The sampling rates for the sub-bands were not changed. This means that energy in the gap regions aliases into the sub-bands. On reconstruction, the gap energy will be restored to its correct position, but will also leave an aliased version within the sub-band. This technique worked remarkably well with no quantization. The aliasing distortion was not very large, and the speech was "fuller" than that with sub-band gaps. However, with quantization the advantage of this scheme disappeared. The quantization noise was more evident than in the previous configuration. One may speculate that this is due to the larger noise bandwidth (and hence noise energy) present. This scheme then trades off fullness for more distortion. The lack of gaps may be important for non-speech applications, where the inability to reproduce tones or spectral components in the gaps may be unacceptable. However for speech, it seems Configuration 2 is slightly more desirable.

5. QUALITY EVALUATION OF RELP AND SUB-BAND

The purpose of the subjective testing was to evaluate the preference of naive subjects for either RELP or SUB-BAND speech at 9.6 kb/s. Formal and informal listening tests in [1] indicate that RELP speech, as produced from the Un and Magill paradigm [4] should produce speech of approximately 4-bit log PCM quality, while SUB-BAND speech should produce a slightly lower quality rating.

5.1 Comparison with Log PCM: Experiment 1

Because of the difference in quality between the RELP and SUB-BAND coded speech, a direct comparison between RELP and SUB-BAND was not carried out. While such a comparison would certainly provide evidence for the preference of one over the other, it would not give an indication of their relative quality with respect to some common scale of signal degradation. The log PCM scale, being familiar to researchers in speech coding, is a convenient scale for such testing purposes. Following [1], it was decided to compare the RELP and SUB-BAND coded speech against log PCM coded speech. Four sample sentences from the Harvard lists of phonetically balanced sentences [21] were recorded by 5 male and 4 female talkers. Four of these talkers (two male and two female) were selected for the subjective testing; all were native speakers of English. Three sentences were selected for testing, as listed in Table 5-1.

TABLE 5-1
Sentences Used in Experiment 1

1. Glue to the sheet to the dark blue background
2. It's easy to tell the depth of a well
3. The pipe began to rust while new

These twelve utterances (3 sentences x 4 talkers) were processed by the RELP and SUB-BAND programs to simulate transmission at 9.6 kb/s. For this experiment, RELP(1) and SUB-BAND (Configuration 2) stimuli were used. Testing of the RELP(2) stimuli is described in Experiment 3 below). Log PCM stimuli at 3, 4, 5 and 6 bit log PCM were created for purposes of comparison. (These stimuli will be denoted by the labels P24, P32, P40 and P48 respectively, indicating 24 kb/s, 32 kb/s, etc.). The stimuli were presented in AB format, where A was either RELP or SUB-BAND and B was one of the four log-PCM stimuli. Each of the selected 12 sentences was coded by the RELP(1) (Section 3.2.4) and SUB-BAND (Section 4.3.2) coders. The coded and uncoded versions of an utterance were matched for loudness so that loudness could not become a cue for preference. Order was fully counterbalanced (i.e., every AB presentation had a corresponding BA presentation). The two stimuli of each diad were separated by a 1 second silent interval, and 4 seconds were allowed after each diad during which the subject could indicate his preference on a score sheet. The factors involved in the test were ORDER (2 orders);

TALKER (4 talkers: MB,MC who were male, FA,FD who were female); UTTERANCE (the three utterances listed in Table 5-1); CODING (RELP or SUB-BAND) and COMPARISON (3, 4, 5 or 6 bit log-PCM), for a total of 192 comparisons.

An audio tape was generated with all factors randomized, and the tape was presented to 10 subjects in two parts, with 96 presentations in each part. The total presentation time was 36 minutes, plus a 10 minute rest period between halves. A total of ten subjects (eight of whom were university students with no previous training as subjects; the remaining two were trained listeners) listened over TDH-39 headphones to stimuli played back using a Revox A77 Mk 4 tape recorder.

Fig. 5.1 shows the percentage of preference judgements along the log-PCM scale for the RELP and SUB-BAND stimuli respectively. Each point represents 24 judgements by each of 10 subjects for a total of 240 judgements per point. The overall results of this test, as determined from the averaged preference curves, that RELP at 9.6 kb/s is equivalent to 4-bit log PCM, while equi-preference on the log PCM scale for SUB-BAND occurs around 3.6 bits.

Figs. 5.2(a) and 5.2(b) show the variability of the preference curves for individual subjects. The spread is observed to be very large. Subject 1, for instance, tended to prefer log PCM speech to either RELP or SUB-BAND speech in virtually all cases except for 3-bit log PCM. Subject 8, on the other hand, evidently preferred both RELP and SUB-BAND over Log PCM. An analysis of these individual preference curves shows that this relative preference of RELP over SUB-BAND speech is consistent across subjects. Thus, while their biases concerning the preference of RELP and SUB-BAND along the log PCM scale vary considerably, the relative ordering is maintained. Table 5-2 lists the estimated equi-preference points on the log PCM scale for each of the subjects, as determined from a Probit analysis in which the preference curves are fitted to a normal ogive [22]. (Subject 6 was omitted since neither the RELP nor SUB-BAND preference curves for this subject reached a value of 50 percent, and since the Probit analysis returned a negative 50 percent intercept on the log PCM scale).

TABLE 5-2 Equi-Preference Values for all Subjects

SUBJECT	N-BIT LOG PCM		
	RELP	SUB-BAND	DIFFERENCE
1	4.7	4.0	0.7
2	3.0	3.3	-0.3
3	3.7	3.2	0.5
4	3.7	3.7	0.0
5	4.9	4.4	0.5
7	3.2	2.8	0.4
8	5.0	5.1	-0.1
9	3.5	3.3	0.2
10	4.7	4.2	0.0
MEAN	4.0	3.8	0.3
STD. ERROR OF MEAN	0.3	0.2	0.1

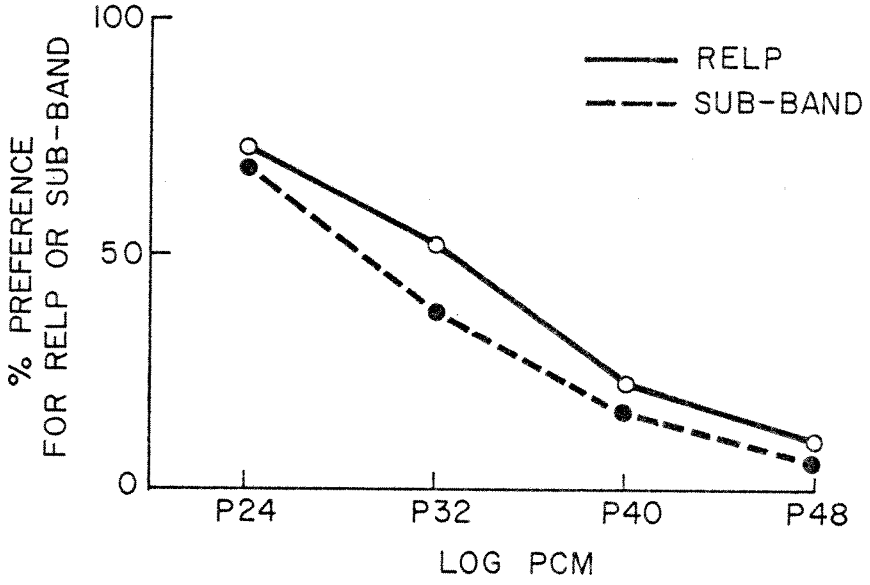

Fig. 5.1 Summary of results of the subjective evaluation of the RELP(1) and SUB-BAND stimuli

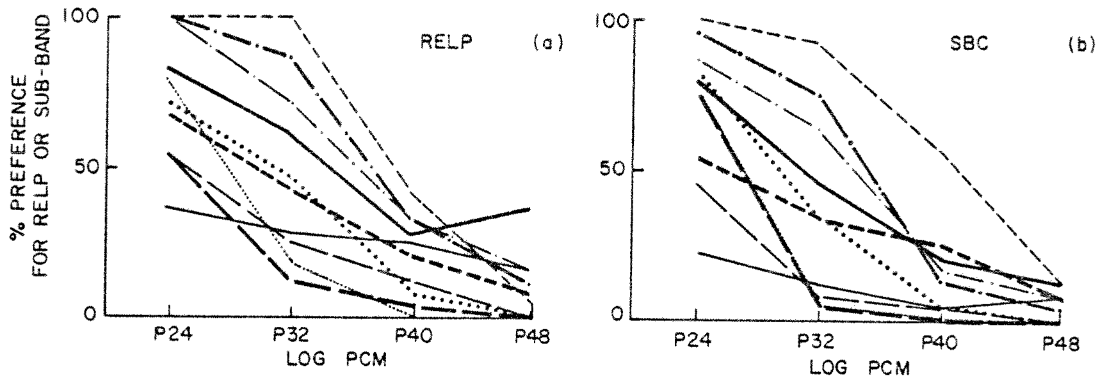

Fig. 5.2 Individual preference curves for the ten subjects. (a) RELP(1) vs. log PCM. (b) SUB-BAND vs. log PCM. Curves belonging to the same subject are indicated by the same line types in (a) and (b).

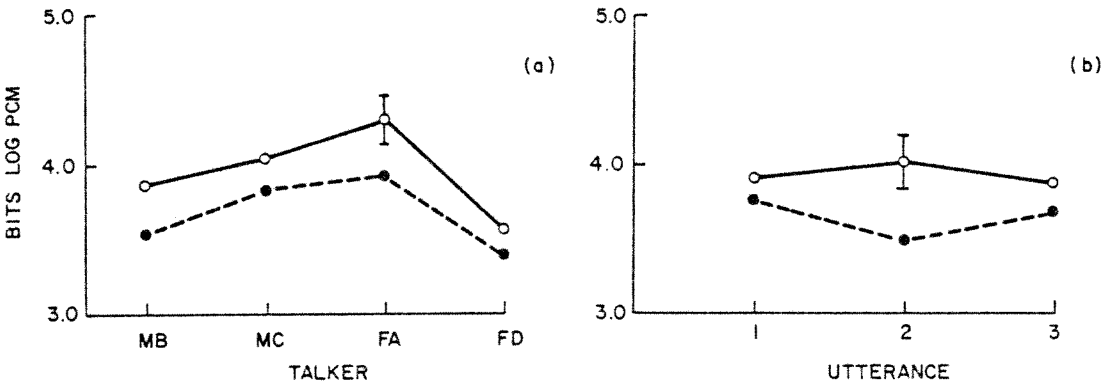

Fig. 5.3 Variation of equi-preference values on log PCM scale for RELP(1) and SUB-BAND stimuli with (a) talker and (b) utterance. In (a), the leading M and F on the labels for the horizontal axis indicate "male" and "female" talkers

It can be seen from Table 5-2 that the preference of RELP over SUB-BAND is maintained for all but two subjects, the average difference being 0.26 bits on the log PCM scale. (This difference may be contrasted with a difference of 0.4 bits obtained from the difference of 50-percent points in Fig. 5.1).

The preference curves for RELP and SUB-BAND, averaged over all subjects, were obtained for each of the factors ORDER, TALKER, and UTTERANCE. The estimated equi-preference log-PCM values for TALKER and UTTERANCE are shown in figs. 5.3(a) and 5.3(b) respectively. A main effect due to TALKER is observed in Fig. 5.3(a), but there does not seem to be any consistent effect due to the sex of the talker. Even so, while considerable variation with talker is observed, the preference of RELP over SUB-BAND is nonetheless maintained; the same is true for the variation with UTTERANCE (Fig. 5.3(b)). The conclusion to be drawn, therefore, is that the subjective assessments concerning the relative placement of RELP and SUB-BAND are robust with respect to TALKER and UTTERANCE. The dominance of the RELP over the SUB-BAND speech is slight, however, and is only of the order of 0.2 - 0.4 bits. Subjectively, the two types of coded speech are characterized by qualitatively different types of signal degradation, and it is difficult to decide which is the more acceptable.

5.2 Comparison with Multiplicative Noise: Experiment 2

To further assess the quality of the RELP and SUB-BAND stimuli and to obtain an estimate of the subjective SNR, a second test was carried out in which RELP(1) and SUB-BAND stimuli were compared against reference stimuli degraded by various amounts of multiplicative noise. Following [23], these comparison signals were produced from

$$s'(n) = s(n) + k\ e(n)\ s(n) \qquad (5-1)$$

where k is a factor less than unity, and e(n) randomly takes on values of either +1 or -1. For this test, a total of 6 sentences were used, representing two talkers (one male and one female) and three separate utterances (those listed in Table 5-1). As in Experiment 1, order was randomized completely. The comparison stimuli represented multiplicative noise levels of 7, 10, 13, 16, 19 and 22 dB. The presentation paradigm was identical to that of Experiment 1, except that the total number of presentations in this case was 144. A total of 7 subjects were tested.

The cumulative results for the seven subjects are shown in Fig. 5.4. A Probit analysis was performed on the preference curves for the individual subjects, and the average of the 50 percent intercepts determined in this fashion was 12.7 for the RELP(1) and 13.2 dB for the SUB-BAND stimuli. Fig. 5.5 shows the relationship between the log PCM scales and the

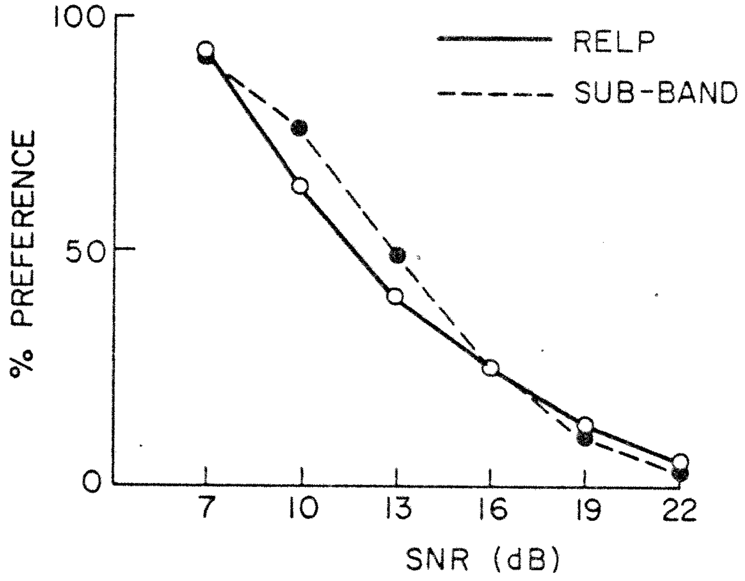

Fig. 5.4 Preference curves for RELP and SUB-BAND when compared against the same utterances with varying amounts of multiplicative noise

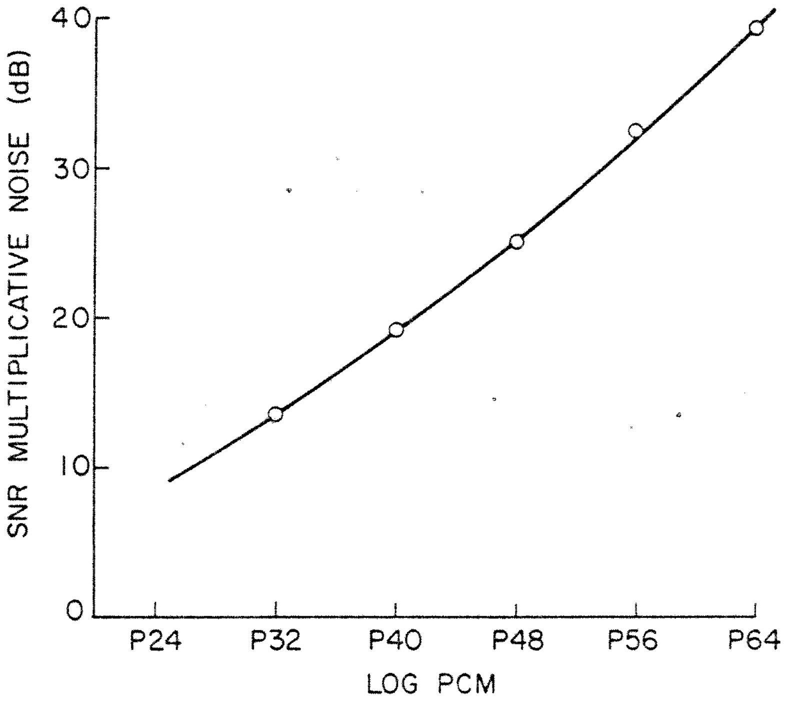

Fig. 5.5 Comparison of log PCM and multiplicative noise scales (from [23])

multiplicative noise scales, as determined by [23]. In this figure it can be observed that the 4 bit log PCM equivalent observed in Experiment 1 is equivalent to an SNR of approximately 13 dB, a result consistent with the outcome of Experiment 2.

5.3 RELP(1) and RELP(2) Against Log PCM: Experiment 3

To obtain an estimate of the amount of improvement provided by this extra differencing procedure, a comparison of the previous RELP stimuli (RELP(1)) and RELP(2) were compared against the same log PCM stimuli. In addition, the SUB-BAND stimuli were included to verify the robustness of the results of Section 5.2, and a 4.5 bit log PCM stimulus was included as a control stimulus. The 4.5 bit log PCM stimulus was created using 23 quantizer levels ($2^{4.5}$=22.63). Thus, the stimulus pairs were always either AB or BA, where A and B are drawn from the lists shown in Table 5-3.

TABLE 5-3
Stimulus Comparisons for Experiment 2

A	B
RELP(1)	3 BIT LOG PCM
RELP(2)	4 BIT LOG PCM
SUB-BAND	5 BIT LOG PCM
4.5 BIT LOG PCM	6 BIT LOG PCM

Four subjects participated in the test, and the results are shown in Fig. 5.6. As in Experiment 1, the RELP(1) and SUB-BAND coded speech returned estimates of approximately 4 bit log PCM, although both are rated slightly lower this time. Fitting the individual preference curves to a normal ogive and averaging the resulting 50-percent intercepts produced the values shown in Table 5-4.

TABLE 5-4
Equi-Preference Values from Experiment 2

CODER	N-BIT LOG PCM
RELP(1)	3.8 ± 0.17
RELP(2)	4.5 ± 0.16
SUB-BAND	3.9 ± 0.19
LOG 4.5 PCM	4.6 ± 0.07

From Table 6-2 it is seen that (a) RELP(1) is rated slightly lower than SUB-BAND and (2), that RELP(2) is approximately 0.7 bits better than RELP(1) on the log PCM scale. Fig. 5.7(a) and 5.7(b) show that this difference is maintained across UTTERANCE and TALKER, at least for the

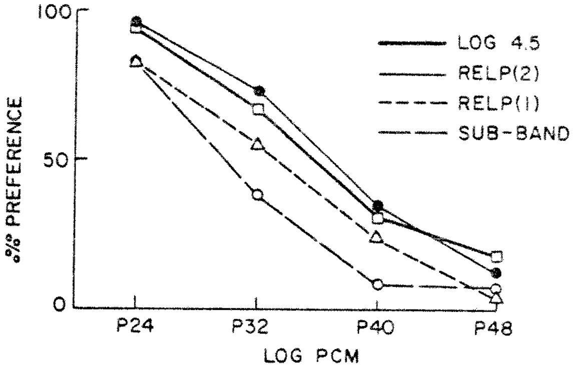

Fig. 5.6 Comparison of RELP(1), RELP(2), SUB-BAND and log 4.5 PCM against the log PCM scale

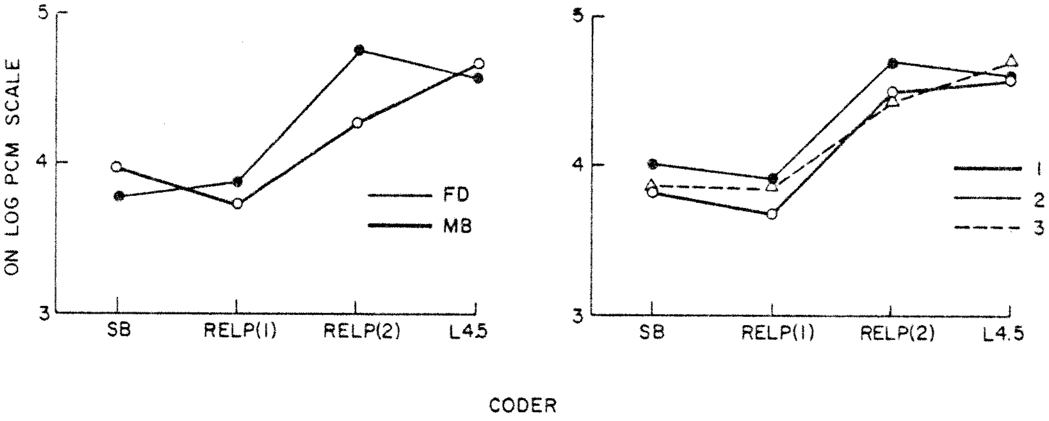

Fig. 5.7 Variation of equi-preference values on log PCM scale for (a) TALKER and (b) UTTERANCE for each of RELP(1), RELP(2), SUB-BAND and log 4.5 PCM

restricted tokens used here. The rating for log 4.5 PCM is the same value (4.6) independently of UTTERANCE and TALKER, as indeed it must be. Significantly, there is also much less variation in the 50 percent estimates (S.E. = 0.07) for the 4.5 bit log PCM comparison.

Both RELP(1) and SUB-BAND were given slightly lower ratings in this experiment, but this is probably just a contextual effect due to the presence of higher quality stimuli (i.e., RELP(2) and log 4.5). An explanation for the high value of 4.6 for the log 4.5 PCM rating is simply that log 4.5 is not "perceptually midway" between log 4.0 and log 5.0.

5.4 Results of the Preference Tests

The results of the preference tests described above suggest that the Un and Magill RELP algorithm (essentially RELP(1)) produces speech which is equivalent in overall acceptability to approximately 4 bit log PCM. Since all of the parameters of the RELP codec were optimized, 4 bit log PCM is considered an upper limit for RELP speech produced by this algorithm. The SUB-BAND stimuli were also rated at 4 bit log PCM equivalent, and it is the opinion of the authors that the RELP(1) and SUB-BAND codecs do indeed produce speech of comparable quality. The modified harmonic generation scheme described in Section 3.2.7 (i.e. RELP(2)), appears to generate speech at or near 4.5 bit log PCM equivalent.

The difference in subjective quality between the RELP(1), RELP(2) and SUB-BAND speech make their absolute placement along the log PCM scale difficult to interpret, but the relative differences appear to be stable. Since the RELP and SUB-BAND ratings were obtained during the same test session, it is reasonable to assume that the relative placement of the test stimuli is an accurate reflection of the listeners' preferences. This conclusion is supported by the fact that the effects of TALKER and UTTERANCE appear to be minimal. The preference for RELP(1) over SUB-BAND (or vice versa) appears to be a strictly personal preference; the cumulative results of Experiments 1, 2 and 3 show that approximately 50 percent of the listeners preferred RELP(1) in favour of SUB-BAND. One hundred percent of the subjects, however, preferred the RELP(2) stimuli over both the RELP(1) stimuli and the SUB-BAND stimuli, independent of utterance and talker.

In a previous study [1], it was concluded that the RELP stimuli were rated as also approximately equivalent to 4-bit log PCM. The results of Experiment 1 show that the RELP(1) stimuli were also rated at approximately 4 bit log PCM equivalent. Fig. 5.8 shows that there is, on the whole, a slight improvement in the RELP(1) stimuli over those of [1]. Fig. 5.8 notwithstanding, it is the opinion of the authors that the RELP(1) stimuli are substantially improved over the previous simulation

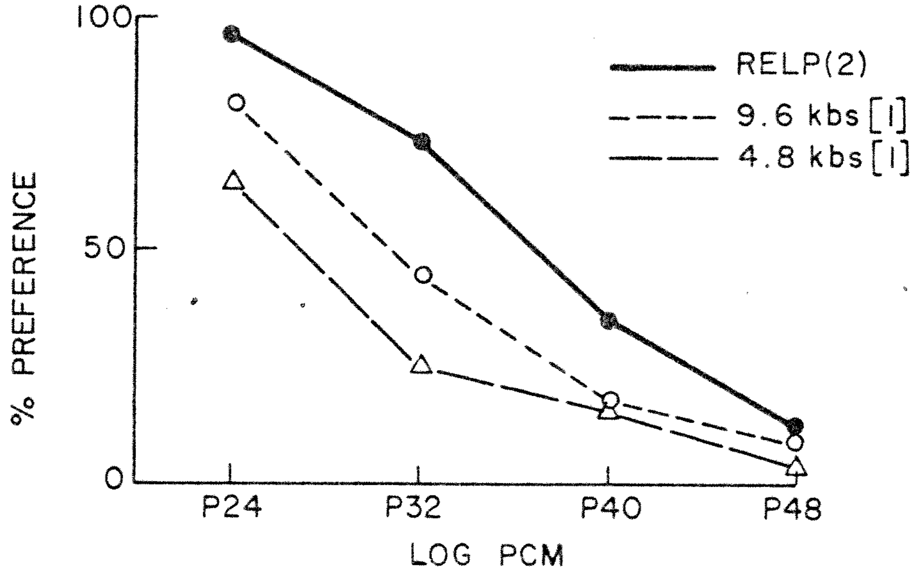

Fig. 5.8 Comparison of preference curves for RELP from Mermelstein et al. [1] and those from the present study

[1]. The RELP(1) stimuli are marked by a single source of degradation - that of a raspy noise-like distortion, in contrast to those of [1] which were dominated by frame synchronous distortions which also interfered with intelligibility. The intelligibility of the RELP and SUB-BAND used in the present study appears very high. The RELP(2) stimuli were marked by less noise than RELP(1), and were thus substantially clearer.

Fig. 5.9 shows assessments of the hypothetical quality of various coders based on the data provided by [1], [24] and the results of the present study. The form of each curve is based on intuitive estimates of coder quality as well as formal subjective estimates using naive subjects. The data symbols in Fig. 5.8 represent points determined by comparison of the designated coder with log PCM stimuli as conducted in Experiment 1 above. The LP (pitch-excited Linear Prediction) curve shows that little additional quality is to be gained by increasing the transmission rate above 2.4 kb/s. Likewise, the RELP curve, assuming a constant baseband of 800 Hz, shows that little improvement is obtained by increasing the bit rate above 9.6 kb/s, but that quality drops off rapidly as the bit rate is decreased below 9.6 kbs. Increasing the width of the baseband would improve the output of the RELP codec, and the subjective quality curve would then asymptote to perfection with increasing bit rate as do the SUB-BAND and log PCM curves in Fig. 5.8. SUB-BAND, according to the present results, is of slightly lower quality at a transmission rate of 9.6 kb/s, but is capable of reaching toll quality (assumed 7-bit log PCM equivalent) at approximately 24 kb/s [24]. The log PCM curve in Fig. 5.9 is anchored by using 7-bit (56 kb/s) to represent toll quality and 4.5 bit (36 kb/s) to represent communications quality [24].

An important consideration in Fig. 5.9 is that the RELP speech at 9.6 kb/s is limited primarily by the method of regeneration of the high-frequency component and not by the quantization of either the spectral parameters or the residual. Thus, to improve RELP speech further, it is necessary to improve the high-frequency regeneration process. Doing so will automatically result in corresponding improvement at 9.6 kb/s. Of course, if the transmitted bandwidth of the baseband residual is allowed to increase as the transmission rate is increased, the subjective quality of the stimuli will also increase, and will asymptotically become distortion free as does the SUB-BAND (Fig. 5.8).

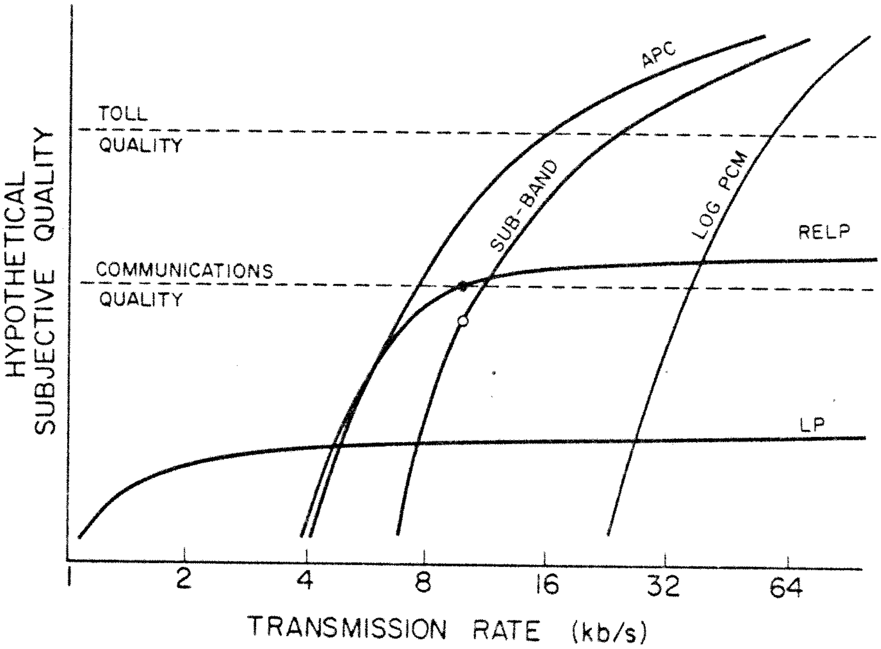

Fig. 5.9 Estimated variation of subjective quality of RELP, PELP (Pitch-Excited LP), SUB-BAND and LOG PCM stimuli as the transmission rate is increased. The RELP curve assumes a fixed baseband of 800 Hz. The data points are (●) RELP and (○) SUB-BAND as determined from Experiment 3 of this report.

6. CODER COMPLEXITY

6.1 RELP Complexity

Figs. 3.1 and 3.2 show the basic structure of the RELP codec. The most complex numerical operation in the encoding stage is that of fitting Equation 3-1 to one frame (typically 160 to 200 points) of the input speech waveform. Assuming that P coefficients are to be extracted, this involves the simultaneous solution of P linear equations, in addition to computation of the autocorrelation functions. The hardware necessary for real-time computation of the a_k (at intervals of approximately 20 milliseconds) is therefore expected to be quite complex. The generation of the residual (Equation 3-4) and the decimation by a factor of (say) 5, as well as the scaling and logarithmic transformation are arithmetic operations which, although adding to the total complexity of the RELP coder, are straightforward to implement.

The output of the RELP coder consists of a bit stream (or, as in the case of the present simulation, an equivalent integer representation). In a real communications link, additional bits would have to be provided for error protection, perhaps up to 5 or 10 percent of the total bit rate, depending on the degree of protection desired. Of course, not all parameters warant protection. The reflection coefficients, for instance, are such that perturbations caused by transmission errors will not generate unstable synthesis filters [10]. However, in the event that protection is deemed necessary, protection of the lower order reflection coefficients is more important than protection of the higher order coefficients.

Recovering the residual from the received signal in the absence of transmission errors is straightforward. In the present scheme, restoration of the absolute gain of the residual is performed at a point prior to low-pass filtering in order to smooth the effects of gain quantization and/or possible errors in transmission of the gain parameter. While it is possible to restore the absolute levels of the residual at various points in the synthesizer, in any practical implementation it would appear preferable to do it at the point suggested above. The various filtering and rectification stages during reconstruction of the residual at the synthesizer can be implemented in such a way that repeated treatment of the whole array of points r(n) is collapsed into possibly a single (but complicated) arithmetic operation. If reconstruction of the residual is carried out on a point-by-point basis, synthesis can then also occur point-by-point.

6.2 SUB-BAND Complexity

The main complexity of the SUB-BAND coder is in the filtering operations. In the simulation, each filter was a 255 tap finite-impulse-response filter. At 9.6 kb/s, little quality would be sacrificed by judicious pruning of the filter lengths to 100 taps. This level of complexity is near the capabilities of Charge Coupled Device transversal filters [25]. This would result in an analog sampled data implementation of the system.

The only truly digital portion of the system would be the quantizers, coders, and multiplexers. The alternative is a digital implementation of the whole system including the filtering operations. The total computational load for the filtering operations is on the order of 2n where n is the filter length and an operation is a multiply-add. For filter lengths in the order of 100, this requires less than a 500 nanosecond multiply-add time. This figure does not include any overhead time for data retrieval. A basic multiplier-accumulator combination capable of these speeds is commercially available as an LSI device (single quantity price about $400) [26]. However this speed is taxing for even a high speed bit slice microprocessor. The only hope for a simple all-digital configuration is a peripheral signal processing chip which can be allied to a basic microprocessor. For now, it seems an all digital version of SUB-BAND coding requires considerable special purpose hardware.

Another study in progress is investigating quadrature mirror filters as a possible answer to the filtering problem [9],[27]. With these filters, a reduced number of taps might be suitable. However, the real advantage of such a scheme is modularity. Each sub-band filter is an exact replica of the others.

7. FUTURE WORK

The results of this study suggest that the upper limit of RELP speech via the modified Un and Magill [4] method of reconstruction of the residual is approximately equivalent to 4.5 bit log PCM, which is generally considered communications quality [24]. The speech produced from this coder is highly intelligible and preserves the identity of the speaker, but suffers from a slight raspiness and hollowness which would limit such a coder to special applications where a low transmission rate was the overriding concern. As is the case with the SUB-BAND coder, achieving better than communications quality RELP speech at 9.6 kb/s is likely only to be accomplished at the expense of a substantial increase in complexity.

A virtue of the RELP and SUB-BAND codecs is that they depend only modestly on the fact that the signal being coded is speech. That is, the operations performed on the signal do not require that any differentiation be made, even on a frame for frame basis, of spectral and/or temporal features of the signal or of its phonetic content. This makes the RELP coder fairly robust against corruption by noise or other background signals. Where the input signal is known to be speech free of interference, it may be possible to carry out different types of high frequency regeneration on a frame-by-frame basis in order to avoid using a harmonic generation scheme for unvoiced sounds [12].

Attempts to improve RELP-coded speech, or for that matter any other type of coded speech, have primarily been carried out in a synthetic fashion: the coding algorithm is altered by changing the value of some parameter, and the effect of the change is then assessed in formal or informal listening tests. In general, little effort has been spent in understanding which type of signal degradations are the least acceptable to listeners, and which are the most acceptable. This information is important since "improving" a coding algorithm for a fixed transmission rate often consists of merely trading one type of signal degradation for another. The RELP(1) and RELP(2) algorithms assessed in this report are good examples of this. Each is characterized by a type of degradation - in the sense that the coded waveform differs from the input signal in a way and by an amount determined by the coding algorithm - but the perceptual effect is much more acceptable in the case of RELP(2). The trade-off in this case is one of a harsh raspiness vs. one of a slight noisiness accompanied by a metallic "sheen". Evidently the second is more acceptable than the first, even though the spectral match to the original is worse (Fig. 3.4). In order to understand why the RELP coder appears to have an upper limit of between 4 and 5 bits log PCM equivalent, it will be necessary to more clearly understand which acoustic characteristics of the RELP speech contribute to the perceived degradation.

The distortions perceived in SUB-BAND coded speech at 9.6 kb/s are the hollowness due to the spectral gaps and the distortions due to the necessarily coarse quantization. Increasing the number of sub-bands beyond the four used in this study will alleviate some of the perceived quantization distortion. However more may be gained if this increase in the number of sub-bands, to say eight, is accompanied by a new dynamic bit assignment strategy. The number of bits assigned to any sub-band would be altered dynamically to better reproduce those sub-bands that have the most energy at the expense of the low level sub-bands. This is indeed the technique used in Adaptive Transform Coding [17]. However a less complex scheme that avoids the linear prediction analysis used in ATC may be possible for SUB-BAND coders.