# Simulation of Digital Coding Techniques for Speech Transmission at 9.6 kb/s

P. Kabal

78-08

INRS-Télécommunications
a/s Recherches Bell-Northern Ltée
3, Place du Commerce
Ile des Soeurs, Qué
H3E 1H6

Rapport technique de l'INRS-Télécommunications no. 78-08

Systems Division

BELL-NORTHERN RESEARCH

TECHNICAL MEMORANDUM

TM 32031 December 1978

# SIMULATION OF DIGITAL CODING TECHNIQUES FOR SPEECH TRANSMISSION AT 9.6 kb/s

78-

PROPRIETARY

Author(s): P. Mermelstein
P. Kabal
D. O'Shaughnessy

Dept.: 3D72
3Q00

recommended by:

approved by:

authorized by:

Keywords : Speech Coding

Abstract : Speech transmission at 9.6 kb/s allows the design of integrated digital networks for speech and data transmission. This report documents progress in the design and evaluation of residual-excited linear-prediction coding and sub-band coding at this rate. Both coders yield highly intelligible speech with quality near that of 4 bit PCM systems.

Funder : Bell Canada
Prime : J.R. Barry
OtherRef : B78404

Case : BR5514

bnr499a

## DISTRIBUTION LIST

Bell Canada

1. *J.R. Barry - Dir. Dev. and Standards, HQTD, 220 Laurier
2. F.M. Banks - Asst. Dir. Dev. and Standards, HQTD, 220 Laurier

Bell-Northern Research

| | | |
|---|---|---|
| 3. *D.A. Chisholm | 0C00 | Corkstown |
| 4. *J.S. Bomba | 9B10 | Central |
| 5. *R.B. Hosking | 3C00 | Toronto |
| 6. P.E. Jackson | 3D00 | Central |
| 7. *G.B. Thompson | 3H00 | Central |
| 8. *A.L. Brosseau | 3B00 | Montreal |
| 9. *R. Kenedi | 3A00 | Central |
| 10.*P.J MacLaren | 3D40 | Central |
| 11.*A.H. Marsh | 3D50 | Central |
| 12.*R.P. Uhlig | 3D20 | Central |
| 13. B. Prasada | 3R10 | Montreal |
| 14. M.L. Blostein | 3R00 | Montreal |
| 15. S. Cohen-Sfetcu | 3Z60 | Central |
| 16. M. Ferguson | 3R30 | Montreal |
| 17. F. Daaboul | 3R20 | Montreal |
| 18. J. Turner | 3R20 | Montreal |
| 19. M. Hunt | 3R20 | Montreal |
| 20. M. Lennig | 3R20 | Montreal |
| 21. V. Gupta | 3R20 | Montreal |
| 22. D. O'Shaughnessy | 3Q00 | Montreal |
| 23. P. Kabal | 3Q00 | Montreal |
| 24. P. Mermelstein | 3R20 | Montreal |
| 25. E. Dubois | 3Q00 | Montreal |
| 26. S. Sabri | 3Q00 | Montreal |
| 27. Tech. Information Center (2) | 8E20 | Central |
| 28. Tech. Information Center (2) | 8E24 | Montreal |
| 29. T.I.C. (1) | 8E20 | Toronto |

* Executive Summary

## SUMMARY

Speech transmission at 9.6 kb/s is of significant interest because that is the highest data transmission rate currently attainable over analog voice lines. Where voice and data are to be transmitted over the same link and connection is in a time-shared mode, 9.6 kb/s is the highest practical transmission rate. To this end two methods of speech coding, residual-excited linear prediction (RELP) and sub-band coding (SBC) are being simulated and evaluated.

Preliminary results indicate that the quality of RELP is slightly more natural than log PCM speech coded at 4 bits/sample. An incomplete study indicates that the quality of 9.6 kb/s SBC is slightly inferior to the same 32 kb/s PCM coding. Although no detailed cost studies have been carried out, the relative complexity of the coding operations involved suggests that sub-band coding is less costly to implement. We need a better understanding of quality/complexity trade-offs of the individual coders to arrive at optimal configuration of the respective techniques before a comparative evaluation can be completed. This work is now being carried out.

We find that we incur a relatively small degradation when the bit rate for the RELP is reduced from 9.6 kb/s to 4.8 kb/s. The currently favoured technique at 4.8 kb/s, linear prediction coding, encounters serious speech quality problems when speech previously passed over a telephone link is encountered. RELP, being more robust, may overcome this problem. Quality comparisons are planned in this area in 1979.

TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## REFERENCES

1. P. Mermelstein (1978), "Evaluation of two 32 kb/s ADPCM Speech Coders for Digital Speech Transmission", BNR TM 32005.

2. P. Noll, B.J. McDermott, R.E. Crochiere, J.M. Tribolet (1978), "A Study of Complexity and Quality of Speech Waveform Coders", IEEE Int. Conf. on Acoustics, Speech and Sig. Proc., Tulsa, OK. 586-590.

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

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

5. J. Makhoul (1975), "Linear Prediction: A Tutorial Review", Proceedings of the IEEE 63, no. 4, pp. 561-580.

6. L. Rabiner, M. Cheng, A. Rosenberg and C. McGonegal (1976), "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Trans. ASSP 24, no. 5 pp. 399-418.

7. M.R. Sambur, A.E. Rosenberg, L.R. Rabiner, and C.A. McGonegal (1977), "On Reducing the Buzz in LPC Synthesis", IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, pp. 401-404.

8. D. Estaban, C. Galand, C. Mauduit, and J. Menez (1978), "9.6/7.2 KBPS Voice Excited Predictive Coder", IEEE Intern. Conf. on ASSP., pp. 307-311.

9. C. Weinstein (1975), "A Linear Prediction Vocoder with Voice Excitation", EASCON, pp. 30A-G.

10. W. Voiers (1978), "Intelligibility Testing at Dynastat: the Diagnostic Rhyme Test", unpublished document.

11. W. Voiers (1972), "Diagnostic Evaluation of Intelligibility in Present-Day Digital Vocoders", Conference on Speech Communication Processing, paper E1, pp. 170-174, IEEE 72 CHO-596-7AE.

12. G.A. Miller and P.E. Nicely (1955), "An Analysis of Perceptual Confusions Among Some English Consonants", J. Acoust. Soc. of America, 27, no. 2, pp. 338-352.

13. A.J. Goldberg and H.L. Schaffer (1975), " A Real-Time Adaptive Predictive Coder using Small Computers", IEEE Trans. on Comm., COM-23 pp. 1443-1452.

14. J. Makhoul and M. Berouti (1978), "Predictive and residual encoding of speech", 96th Mtg. of Acoust. Soc. of America, 64, Supplement 1, page S138.

15. R.E. Crochiere, S.A. Webber, and J.L. Flanagan (1976), "Digital Coding of Speech in Sub-Bands", IEEE Intern. Conference on ASSP, pp. 233-236 (also Bell System Tech. Journal, 58, pp. 1069-1085).

16. B.S. Atal and M.R. Schroeder (1970), "Adaptive Predictive Coding of Speech Signals", Bell System Techn. Journal, 49, pp. 1973-1986.

17. B.S. Atal, M.R. Schroeder, and V. Stover (1975), "Voice-Excited Predictive Coding System for Low Bit-Rate Transmission of Speech", Intern. Conf. on Comm., 2, pp. 30:37-30:40.

18. M.R. Sambur (1978), "High Quality 9.6 kbps Algorithm that Satisfies the Embedded Bit Concept", Intern Conf. on Communications, Vol. I, pp. 12A.2.1-4

19. J.L. Flanagan (1972), Speech Analysis, Synthesis, and Perception, Second Edition (Springer-Verlag: Berlin).

20. N.S. Jayant (1973), "Adaptive Quantization with a One-Word Memory", BSTJ, 55, pp. 1119-1144.

## 1. INTRODUCTION

This report documents results attained up to this time in the simulation and evaluation of two techniques for digital speech transmission, residual-excited linear prediction coding and sub-band coding. The target transmission rate is 9.6 kb/s, appropriate for speech transmission over digital networks, for example an integrated private voice-data network. The quality of speech coded at this rate is generally not as high as the current 64 kb/s PCM standard. Yet the speech is quite intelligible and considered acceptable among special groups of users requiring reduced communication costs.

The objectives of this work are to simulate and evaluate known techniques, understand their limitations and possibly improve their performance. Known speech coding techniques range from simple waveform coding techniques appropriate for high bit rates, such as 64 kb/s mu-255 PCM coding, to complex spectrum coding at 2.4 and 1.2 kb/s such as channel vocoding and linear-prediction coding. The two coding processes considered here approach the 9.6 kb/s rate from opposite directions. Residual-excited linear prediction coding attempts to make linear prediction coding more robust by eliminating the need to extract a pitch rate from which an excitation is regenerated at the receiver. Instead the residual component, the component in the signal remaining after spectrum variations has been eliminated, is directly encoded at some additional cost in increased transmission rate. In contrast, sub-band coding attempts to make waveform coding more efficient by separately encoding the waveform components found in separate frequency bands and thereby reducing the total transmission rate.

Pitch-excited linear prediction codecs include an algorithm to decide whether the speech at any given moment is voiced or unvoiced, i.e., whether vocal cord vibration is present. Where the speech is determined to be voiced, the pitch period has to be determined accurately. The method for regenerating the speech signal depends crucially on this decision and errors in voicing have serious effects on the naturalness of the decoded speech. The tradeoffs between pitch and residual excitation involve the higher transmission rates needed to transmit the residual information, the improved quality of the residual-excited signal and the comparative complexities of the two algorithms. The added complexity of the residual-excited (RELP) codec is estimated at about 2-3 times that of a pitch-excited (PELP) codec due to the extra computations needed to encode the residual.

The intelligibility of RELP speech is comparable to that of PELP speech, while giving speech quality in terms of naturalness superior to that of PELP speech. RELP is also expected to be much less deteriorated when adverse conditions degrade the original speech. PELP requires a pitch detector, which has significant difficulties with noisy or telephone speech, and thus PELP speech suffers with degraded speech input. RELP, since it does not extract pitch explicitly, avoids these problems, and should be more robust to such speech input.

Simple waveform coding techniques (PCM, differential PCM, and ADPCM) have been shown to yield toll-quality (7 bit log PCM) speech down to 32 kb/s [1]. More complex techniques such as transform coding [2] yield very good speech quality at 16 kb/s. Sub-band coding attempts to achieve a compromise by preserving as much as possible the simplicity of the waveform coding techniques yet minimizing the reduction in speech quality due to a reduced transmission rate. In sub-band coding several independent quantizers are used to encode the speech signal in a number of contiguous frequency bands. The method takes advantage of the different criteria of the human listener to quantization noise in the different frequency bands. Thus more bits are assigned to transmit the lower frequencies than the higher frequencies. Determining the upper and lower cutoff frequencies of the several frequency bands and deciding on the number of bits to allocate to each are just some of the questions that require detailed examination. At 16 kb/s sub-band coding has been found to yield speech quality comparable to 5 bit log PCM (40 kb/s PCM speech). Designs for best quality at 9.6 kb/s are under investigation.

This is an interim progress report. No recommendations can be made as yet as to the preferred technique for any specific application. In particular, further work is required to establish the specific factors that represent the main impediments to improved speech quality for each coding technique.

## 2. CONCLUSIONS

Separate listening tests were performed to estimate the intelligibility and naturalness of the RELP speech. The intelligibility test required indentification of the initial consonant in isolated monosyllabic words. Each of five subjects listened to 232 words processed by a simulated 9.6 kb/s system. The results showed 92.5% correct word identification. This result is comparable to the best intelligibility figures on PELP speech at 2.4 kb/s. Thus intelligibility is maintained by use of residual coding instead of pitch-excited coding.

A second test showed RELP speech at 9.6 kb/s to be slightly more natural than log PCM speech coded at 4 bits/sample. At the same time we tested a 4.8 kb/s version of the RELP and the quality was found to be just slightly worse, roughly midway between 3 and 4 bits/sample log PCM. While the quality of the RELP speech is considered better than that of PELP, detailed tests have not yet been carried out. We expect to be able to improve the naturalness of the RELP coded speech with the aid of additional research. The evaluation cited only indicates the quality attainable by direct simulation of the technique as described by Un and Magill [3].

The main problem with the reconstructed speech is the "harsh" voice quality of the vowels which sounds as if the speaker had a breathy voice. This can be attributed to the inaccurate reconstruction of the high-frequency component of the residual. Therefore more accurate reconstruction of the waveform of the residual, possibly by some waveform coding technique, represents the first goal in attempting to achieve improvements in the speech quality of RELP coding.

For sub-band coding we find that we can achieve good quality speech coding at a 16 kb/s transmission rate. The coder is robust in the sense that the quality of the reproduced speech does not significantly depend on the speaker, e.g., male or female. This is in constrast to linear prediction coding which does demand high-quality input speech and is somewhat speaker-dependent.

The promise of sub-band coding at 9.6 kb/s is still under investigation. The present view is that it should be possible to design a 9.6 kb/s coder that is only slightly inferior in quality to the 16 kb/s coder demonstrated.

According to current estimates, both the residual LPC and the sub-band coding should yield qualities comparable to 4 bits/sample PCM, the RELP slightly better than the sub-band. On the other hand, the sub-band appears simpler and less costly to implement. We expect to define the cost/quality tradeoffs better in future work.

## 3. RESIDUAL-EXCITED LINEAR PREDICTION

### 3.1.0 Theory

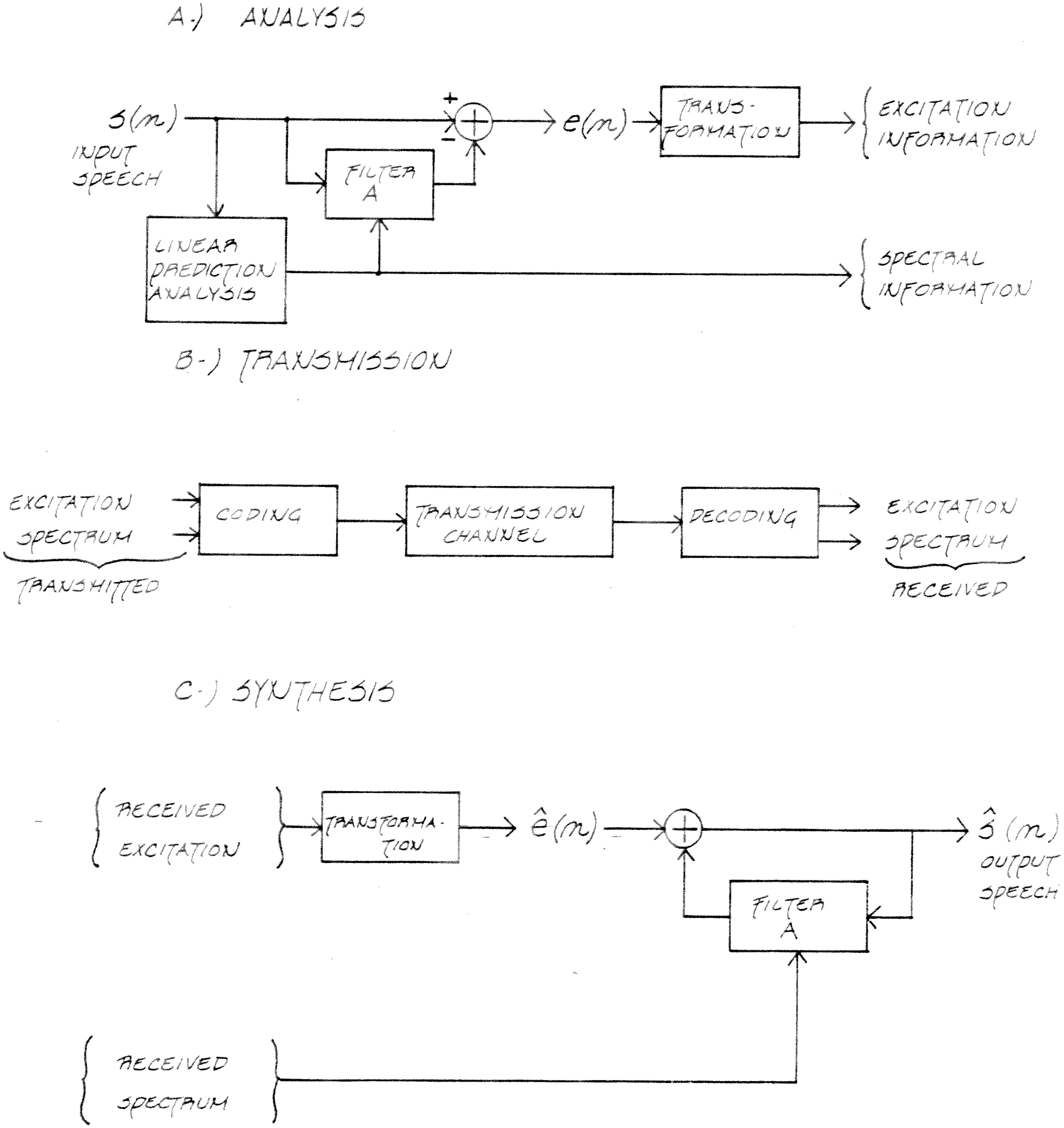
This section on the theory of a residual LP codec covers the main points of linear prediction for speech, ranging from the LP analysis to the transmission of speech parameters and the synthesis-reconstruction of the speech. The differences between pitch-excited and residual-excited LP are established, and the advantages of each are discussed.

### 3.1.1 Generalized Linear Prediction Model

Speech codecs (or vocoders) that employ linear prediction all use the basic model shown in Fig. 1. An LP analysis is performed on the incoming speech s(n) in sections (called windows) of approximately 15-40 msec each, resulting in a set of p parameters known as LP coefficients ($a_i$, i=1,...p). These parameters contain the primary information about the spectral content of the speech during a given window. This process is repeated every T seconds, where 1/T is known as the frame rate; typical values for T are in the 10-30 msec range. The set of LP coefficients, updated every T seconds, determines a variable predictor filter A, which estimates s(n) based on the p previous values of s(n):

$$s(n) = \sum_{n=1}^{p} a_i \; s(n-i)$$

The LP process is based on the assumption that most speech can be well modeled as an all-pole process with a small number of poles (i.e., p poles) and thus knowledge of the prior p samples of s(n) allows a good estimate of the next sample [4,5]. This approximation to the actual physical nature of speech becomes better as p is increased and T is decreased, but good modeling is possible at T = 20 msec and p = 10 for speech containing no frequencies above 4 KHz. Since the transmission bit rate of the spectral information varies as p/T (because p coefficients must be transmitted every T seconds), there is a direct tradeoff of bit rate versus good LP modeling (and hence good quality speech with the LP model). T must be small enough to capture sudden transitions in the spectral information, which correspond to rapid movements of the speaker's vocal articulators (e.g., jaw, tongue, lips). Since some rapid speech sounds are as short as 5 msec, T should ideally be on the order of 5 msec. However, the vast majority of speech sounds are longer than 20 msec, so that value was chosen for this study. This problem can be partially avoided using more complex LP codecs known as variable frame rate vocoders, which transmit spectral frames only when significant changes have occurred in the LP spectral representation [4]. These algorithms were estimated to be beyond the scope of this study, but certainly should be considered in any actual development of a RELP codec.



A.) ANALYSIS

B.) TRANSMISSION

C.) SYNTHESIS

Figure 1 - Overall System for a Linear Prediction Vocoder

The accuracy of the LP representation increases with p, the number of poles in the model. If p is sufficiently large, a number of poles will exist in complex conjugate pairs and correspond to spectral peaks or "formants" in the spectrum of the windowed speech signal. These formants are the main carriers of spectral information, and occur on an average of one formant per 1 KHz of spectrum for an adult male (for people with shorter vocal tracts, e.g., women and children, the ratio is lower). Thus, when dealing with speech low-passed to 4 KHz (as was the case in this study), the model should have at least 8 poles corresponding to the 4 formants expected. Usually extra poles are included to correspond to the spectral shaping that occurs at the glottis. In this study, p = 8 and 12 were used at 4.8 and 9.6 kb/s rates, respectively.

The LP coefficients ($a_i$) are determined by examining a window of speech of length N samples (here, N = 204, corresponding to 25.5 msec at 8000 samples/s), and minimizing

$$\sum_{n=1}^{N} (s(n) - \sum_{i=1}^{p} a_i \, s(n-i))^2$$

These $a_i$ provide the minimum mean square error for the LP model. Various algorithms are available to efficiently evaluate the $a_i$, e.g., autocorrelation, covariance, and lattice methods [5].

### 3.1.2 Linear Prediction Analyzer/Coder

The first stage of an LPC system involves the transformation of the speech signal into spectral and excitation information. This section examines the analysis procedure, and describes what form the output data takes.

Fig. 1 shows that the output of the predictor filter A (i.e., the estimate of s(n)) is subtracted from the actual s(n) to yield an error, or residual, signal e(n). The size of e(n) depends on the accuracy of the LP model. If at time n the speech is well modeled by p poles and the spectrum is not changing rapidly, e(n) should be small in magnitude compared to s(n). Assuming adequate values for T and p, there still remains the problem of finding the excitation for the p-pole process that the LP model provides. The LP algorithm assumes no excitation over the window of speech it is modeling; thus when in fact there is external excitation, as when the vocal tract is excited by turbulence noise at a constriction somewhere along its length (unvoiced speech) or when bursts of air emerge from an oscillating set of vocal cords (voiced speech), one can expect e(n) to be larger. If e(n) is consistently large, unvoiced speech is suspected; whereas if e(n) is low but has quasi-periodic sudden bursts of amplitude, voiced speech is likely, and the time between bursts is estimated to be the fundamental period. The inverse of the period, called the fundamental frequency or pitch, is an estimate of the rate of vibration of the vocal cords, i.e., the rate at which the vocal tract is excited with bursts of air.

In PELP codecs, e(n) is not transmitted to the decoder-synthesizer; indeed e(n) need not be calculated at all (eliminating the need for predictor filter A). Instead, once per frame, a voiced/unvoiced decision and a pitch period estimate (if voiced) is sent. This reduces considerably the transmission bit rate, but requires the added complexity of an algorithm (called a pitch detector) to determine these values. The accuracy of these pitch detectors is crucial to the performance of the codec, and thus there exist many algorithms and implementations both in software and hardware [6]. Unfortunately, while most pitch detectors give good performance most of the time, they also have errors: some minor (e.g., slightly mis-estimating the pitch period) and some major (grossly misjudging the period, or making the wrong voiced/unvoiced decision). Small errors in pitch period are not very relevant perceptually in terms of the synthesized speech output, but gross mistakes can have more severe perceptual consequences. Particularly annoying to people listening to speech emerging from a PELP codec are pitch detector errors which cause voiced speech to be pronounced as if whispered (a voiced-to-unvoiced error), and errors which cause insertion of extraneous voiced periods (unvoiced-to-voiced errors). Such errors are more commonplace for speakers with extreme pitch ranges (e.g., baritones, sopranos) and for deteriorated speech (e.g., noisy, or telephone speech).

Assuming one could devise a perfect pitch detector, there still would remain the problem of the unnatural quality of the PELP speech. During voiced intervals, such speech has a distinct "machine-like" quality and sounds excessively "buzzy", which is primarily due to the reduction of information in e(n) from a complex waveform sampled 8000 times a second to a single pitch estimate at the frame rate (e.g., 50/s).

### 3.1.3 Decoder/Synthesizer

The second stage in an LP system concerns the reconstruction of the speech from information received over a transmission channel from the analysis/coder. This section briefly looks at some problems that occur in attempting to reconstruct speech from a representation with less information than in the original speech.

The problem arises at the decoder as to how to reconstruct e(n), so as to drive an inverse filter 1/(1-A), and recover the original speech. If e(n) and $a_i$ are transmitted directly with no coding, then the output speech $\hat{s}(n)$ would be exactly the same as the input s(n). However, in practical systems, both are transformed and quantized, and thus the versions of e(n) and $a_i$ received at the decoder (termed $\hat{e}(n)$ and $\hat{a}_i$) exhibit some degradation compared to the original values.

As noted above, PELP codecs transmit only a pitch estimate, and thus many such codecs have $\hat{e}(n)$ consist of either white noise (simulated by a random number generator, to correspond to e(n) in unvoiced speech) or a series of impulses (with the pitch estimate specifying the duration between impulses, for voiced speech). This is a reasonable model for e(n) and gives good perceptual results for unvoiced speech, but produces unnatural yet intelligible voiced speech. The fine structure of e(n) in voiced speech which represents the imperfections in the LP modeling is lost when a mere pitch estimate replaces the waveform e(n). It is this lack of fine structure which gives the PELP speech its buzzy quality. Several atempts have been made to ameliorate the unnaturalness, with limited success [7].

In summary, PELP codecs give intelligible but unnatural speech at low bit rates near 2.4 kb/s. RELP codecs, as described below, yield more natural -sounding speech with no loss of intelligibility at rates of 4.8-9.6 kb/s.

### 3.1.4 RELP Coding compared to Pitch-Excited LP Coding

This section describes how RELP codecs differ from PELP codecs, and how their speech improves upon PELP quality.

RELP codecs actually transmit a transformed version of e(n) to the receiver and thus retain much of its fine structure, and avoid completely the buzzy characteristic of PELP speech. Unfortunately, it is not possible to directly quantize and transmit e(n) to the decoder at these bit rates. For example, with the 8000/s sampling rate used in this study for speech low-passed to 4 KHz, a 4 bit log PCM transmission of e(n) would result in a bit rate for e(n) alone of 32 kb/s. To achieve 9.6 kb/s or less for e(n) and $a_i$, the information in e(n) must be reduced in some manner.

There are currently two basic types of RELP codecs, both of which utilize the same fundamental principles. One transmits a low-passed version of e(n) to the receiver, and will be referred to as the residual-transmitted RELP (RRELP) codec [3,8]. The other transmits a baseband of s(n) (called b(n)) to the decoder, where a transformation on the received baseband results in $\hat{e}(n)$. This latter will be called the baseband-transmitted (BRELP) codec [9].

Thus in both cases a low-pass signal is transmitted. Typically, e(n) or s(n), which contain frequencies up to 4 KHz, is passed through a digital low-pass filter with a cutoff of 800 Hz, and decimated 5:1 to result in a fivefold decrease in the bit rate for the residual or baseband.

The problem at the receiver remains to reconstruct e(n) adequately from a low-passed and quantized $\hat{e}(n)$ or $\hat{b}(n)$. Better results are obtained in the RELP than PELP codecs since more information is present in the former. In both RELP codecs, 1:5 interpolation and low-pass filtering, to get the signal back to 8000 samples/s, is followed by a non-linear distortion operation on the received excitation signal. This non-linear operation restores some energy at higher frequencies while retaining the proper harmonic structure present in the low-pass signal. One such operation is a full- or half-wave rectifier. Typically, it does not restore sufficient energy at the higher frequencies, and one of two methods are used to boost the gain there. The RRELP codec uses a double-differencing operation to give a +12 dB/octave gain above the 800 Hz cutoff. The BRELP cannot use so simple a device because the output would not sufficiently resemble e(n). Instead, an LP analysis and predictor filter operation is performed on $\hat{b}(n)$ to result in $\hat{e}(n)$. Thus both RELP codecs perform an LP analysis and predictor filtering to yield a residual signal, but the RRELP does it in the coder, while the BRELP does it in the decoder. The coder complexity is greater for RRELP than BRELP, but the corresponding decoder is simpler. Actually, the BRELP requires two LP analyses, one at the coder to determine the $a_i$ and one at the decoder to obtain $\hat{e}(n)$. Since a double-difference operation is much simpler than an LP analysis, the BRELP is computationally more complex.

The BRELP has a theoretical advantage in bit rate transmission in that the number of poles used in the LP analysis model for the decoder can be arbitrarily large (and hence the modeling of $\hat{b}(n)$ can be very good), since the resulting spectral coefficients are not transmitted, but merely used to drive the predictor filter. However, while the number of poles in the decoder LP analysis does not affect the bit rate, it increases the decoder complexity.

The BRELP codec uses the output of the predictor filter as $\hat{e}(n)$, to drive the synthesizer. The RRELP, on the other hand, passes the output of the double-differencer through a digital high-pass filter (with cutoff 800 Hz). Then the low-pass $\hat{e}(n)$ as received and interpolated by the decoder is added to the high-pass version of the reconstructed $\hat{e}(n)$ to yield the full estimate of e(n) to drive the synthesizer. In the case of RRELP, the synthesizer output is called $\hat{s}(n)$, the final reconstructed speech; whereas the BRELP passes its synthesizer output through a digital high-pass filter, and adds that output to $\hat{b}(n)$ to yield $\hat{s}(n)$.

Thus the BRELP takes advantage of the existence of an accurate (albeit quantized) version of the baseband at the decoder, and uses that signal as the baseband of the output $\hat{s}(n)$, relegating the effects of the LP model and reconstruction to the higher frequencies. The RRELP instead takes advantage of an accurate low-pass $\hat{e}(n)$ at the decoder and uses that as the baseband portion of the residual to drive the synthesizer.

If a high-pass filter is not used in the decoder to enable the inclusion of the received $\hat{e}(n)$ or $\hat{b}(n)$ AFTER the distortion process, the complexity of the decoder is decreased, but the output speech is also substantially degraded (with no change in the transmission rate). Thus the extra complexity is necessary.

### 3.2.0 Simulation

The RRELP and BRELP codecs were implemented in Fortran software on the PDP-11/45 at BNR Montreal, and tested on many samples of speech from both men and women. This chapter deals with specific details of the implementations.

### 3.2.1 Analysis

Speech was recorded in a quiet room onto audio tape at 7.5 ips, and then transferred to computer disk and digital tape storage via a 15 bit A/D converter sampling at 8000/s. The A/D was preceded by an analog low-pass filter with cutoff set just under 4 KHz. The speech was then processed by software simulating the RELP codecs at various transmission rates. The reconstructed speech was moved to audio tape (again 7.5 ips) via a 15 bit D/A converter followed by an analog low-pass filter with 3.8 KHz cutoff. These audio tapes were then used for perceptual experiments. Prior to the actual RELP simulation, the digitized speech was preprocessed with a simple pre-emphasis algorithm ($y(n) = x(n) - .95 x(n-1)$, where $x$ = input speech and $y$ = pre-emphasized speech), which gave a +6 dB/octave boost to frequencies above about 65 Hz. This aids the LP modeling by reducing the dynamic range of the speech spectrum, which normally has a fall-off of -6 dB/octave. A corresponding post-processor de-emphasized the output speech after the RELP simulation.

The RELP simulation used a frame rate of 50/s, and a 25.5 msec Hamming window for analysis. The autocorrelation LP analysis modeled 8 poles for the 4.8 kb/s speech and 12 poles for the 9.6 kb/s speech. A direct form inverse filter provided e(n) for the RRELP. A 31-tap digital low-pass filter with cutoff at 800 Hz was used to generate the low-pass residual or baseband for transmission. The filter response was flat within 1 dB of unity gain up to 400 Hz, and then monotonically fell to -5 dB at 600 Hz, and finally to -34 dB at 840 Hz; above 840 Hz, the response remained below -34 dB. Thus there was very little energy above 800 Hz, which allowed a 5:1 decimation of the sampling rate from 8000/s to 1600/s.

### 3.2.2 Coding/Transmission

Various coding techniques were explored for the transmission of the low-pass residual or baseband, as well as for the spectral information. e(n) has a relatively-flat spectrum because it results from an inverse filtering operation on s(n), i.e., the 1-A filter removes most spectral content such as formants from s(n), leaving harmonic structure but with flattened amplitude. As a result, differential or adaptive quantization schemes yielded little improvement over straightforward log PCM (Pulse-Code Modulation). An ADM (Adaptive Delta-Modulation) scheme, transmitting one bit of information at a rate higher than 1600/s was also attempted, but this too gave results similar to log PCM. So log PCM was used to transmit e(n): 2 bit/sample for the 4.8 kb/s speech and 4 bit/sample for the 9.6 kb/s speech. Thus 3.2 and 6.4 kb/s were used to transmit the residual in the respective codecs, leaving 1.6 and 3.2 kb/s for the spectral information.

Rather than transmit the p LP spectral coefficients directly to the decoder, the $a_i$ were transformed to another set of p parameters called reflection coefficients ($k_i$). The $k_i$ have more desirable transmission characteristics than the $a_i$: they have a smaller dynamic range (the $k_i$ are restricted to -1 to +1, while the $a_i$ are not suitably restricted), guarantee stable synthesis filters (while the $a_i$, after possible errors in transmission, do not), and do not require as much accuracy for transmission (i.e., equal changes in $a_i$ and $k_i$ result in bigger spectral differences for the $a_i$ than the $k_i$). Thus the $k_i$ were transmitted at a rate of 50 frames/s, allowing 64 and 128 bits/frame for the 4.8 and 9.6 kb/s speech codecs. Included with the spectral information in being transmitted once per frame were 2 gain paramters: one reflecting the average energy in the transmitted residual or baseband signal, and one (only used for the RRELP) indicating the "normalized residual" (i.e., the energy ratio of the residual to the total signal). Thus, for the 4.8 kb/s speech, 9-10 parameters (8 reflection coefficients and 1-2 gain parameters) were transmitted in 64 bits, and for the 9.6 kb/s speech, 13-14 parameters (12 reflection coefficients and 1-2 gains) in 128 bits. The reflection coefficients are obtained in a manner such that most of the spectral information resides in the low-order $k_i$; so $k_0$ was assigned the most bits and $k_p$ the least, and the gain parameters were assigned an average number of bits.

Because most spectral changes occur over periods of time greater than 20 msec, individual $k_i$ often exhibit significant frame-to-frame correlation. Thus a differential PCM coding scheme was used for the $k_i$ transmission. A perfect transmission channel was assumed (i.e., no transmission errors), so the effects of transmission were limited to those of quantization.

### 3.2.3 Synthesis

At the decoder, the received $\hat{e}(n)$ or $\hat{b}(n)$ was interpolated 1:5 and passed through a digital low-pass filter, identical to the one in the coder (Fig. 2). In the RRELP, $\hat{e}(n)$ was then full-wave rectified, sent through a double-differencer with an 800 Hz breakpoint, and processed by a digital high-pass filter with 31 taps and an 800 Hz cutoff. Specifically, the high-pass filter was the inverse of the low-pass filter: frequency response was below -34 dB until 400 Hz, rose monotonically to 0 dB at 840 Hz (passing through -6 dB at 600 Hz), and remained within 1 dB of unity gain above 840 Hz. This high-pass signal was then added to $\hat{e}(n)$, with appropriate weighting given to each to balance the high/low frequency energy distribution in the original $e(n)$. This signal was then added to white noise (generated by a Fortran IV random number algorithm). These two signals were weighted according to the transmitted gain parameters: the overall gain was specified to match that of the original $e(n)$, and the amount of added noise was determined by the normalized error with more noise contributed if $e(n)$ constituted a large part of $s(n)$ (corresponding to unvoiced speech). For more details, see reference [3], which this simulation followed closely except in the coding procedures and choice of digital filters.

### 3.3.0 Experimental Results

Two perceptual experiments were performed using the RELP simulation described in the previous chapter. Section 3.3.1 describes an experiment to ascertain how intelligible the RELP speech was, while 3.3.2 deals with explanations of causes for the perceptual errors in the experiment. Section 3.3.3 notes an experiment to compare RELP speech with log PCM speech for naturalness.

### 3.3.1 Intelligibility Test

232 monosyllabic words spoken in isolation by an adult male were processed using the 9.6 kb/s RRELP simulation, and recorded in random order on a test audio tape interspaced with short pauses. The original speech was provided on a tape from Dynastat, the company that designed the Diagnostic Rhyme Test (DRT) used in this intelligibility test [10,11]. The basic thrust of the test is to determine how well phonemic features crucial to English speech perception are preserved after the speech has been processed. Specifically, the features tested are: voicing, nasality, sustention, sibilation, graveness, and compactness. These are all consonant features, and the test makes no attempt to identify vowel confusions. An assumption seems to be that testing for consonant intelligibility is adequate, and that separate testing on vowel intelligibility is unnecessary. Possible reasons for this include: the easier confusion of consonants than vowels (24 consonants versus 11 vowels and 4 diphthongs), the specification of vowels in steady-state by proper location of formants rather than the more complex and transient consonant cues (with corresponding greater difficulties in replicating consonants than vowels in most speech processing algorithms), the perception of consonants categorically (i.e., small acoustic changes can

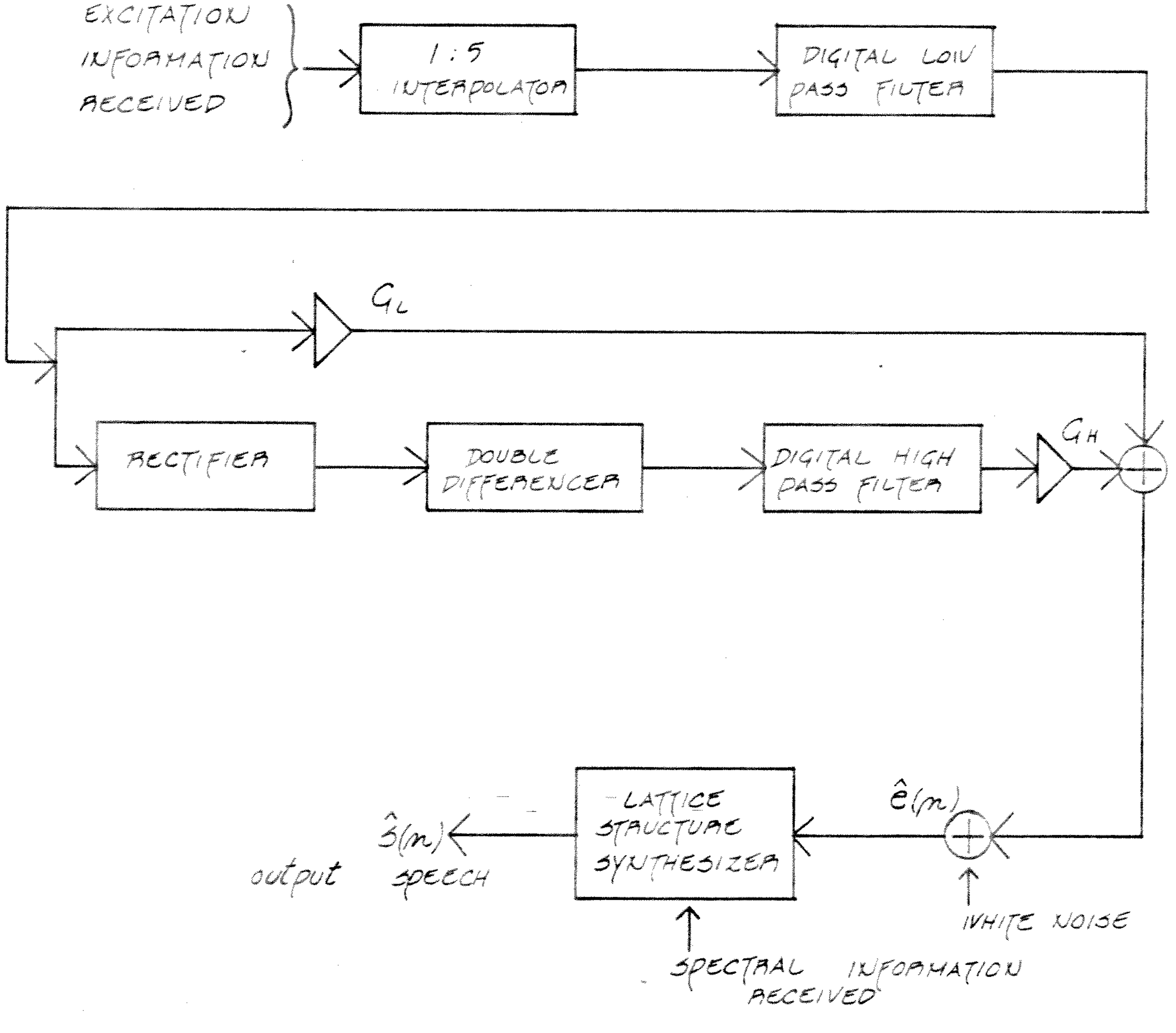


Figure 2 - Decoder for a Residual-Excited Vocoder

lead to large perceptual changes) while vowels are perceived continuously (varying the formants gradually changes one vowel to another), and the greater agreement among listeners on consonant perception than vowel perception (e.g., 2 listeners might interpret the same utterance as "bit" or "beat", but given a choice between "bit" and "pit", they would not disagree).

The form of the test exploits this phenomenon in that a processed word is played through earphones to the listener, and he is asked to choose between 2 word candidates written on the answer sheet. (The words were orthographically familiar to the subjects, and not "nonsense" syllables.)

The original spoken word corresponds to one in the written word pair, and the other is phonemically identical to the first, except for one feature in the initial consonant. By using only isolated words, and controlling all but one feature in the word pair, the DRT presents the listener with a perceptual task significantly greater than that in normal speech. Usually, speech is uttered in sentences and the listener has a good idea of the context in which the speech occurs. Thus slurred or missing acoustic cues can be compensated by the listener's expectations of what is to be spoken. Hence one expects a lower performance on the DRT than on an intelligibility test with sections of speech longer than monosyllables. An error rate of 7.5% (as obtained in this study) should not be interpreted to mean that listeners would mistake one of every 13 phonemes in normal speech, since contextual cues would correct most errors. The DRT's main value lies in ascertaining an upper bound on the intelligibility errors.

The 232 words were presented to 5 subjects, yielding 1160 trials, in which there were 87 errors. As a control, the original tape was also played for 4 of the listeners, who made a total of 8 errors. This gives a rough indication of the difficulty of the test (this control performance would likely be worse if, instead of using the original tape, the computer pre- and post-processing is included; i.e., separate from the RELP algorithm, the aspects of A/D and D/A conversion with appropriate low-pass filtering can be expected to degrade the speech slightly).

### 3.3.2 Detailed Examination of Errors

a) Voicing: In 160 trials, 4 errors were made (2.5%). This category tested pairs such as "best-pest" and "jock-chock" (where the first in each pair has the voicing feature and the second does not). Voicing is a phonemic feature only for the 6 stops, 2 affricates, and 8 fricatives (other than /h/) in English. As was the case in all word pairs, all other aspects of the words were the same: the vowel and final consonants were identical, and the initial consonant (no clusters were used) preserved the same manner and place of articulation. The 100 trials with stops and affricates led to no voicing errors, since the 4 errors all occurred in the 60 trials with fricatives. Since voicing in stops is often indicated via timing information, while in fricatives it is the presence of a low-frequency voice-bar, this perceptual disparity suggests good timing preservation by the RELP, but not as good voice-bar representation.

b) Nasality: 2 errors were made in 160 trials (1.25%), testing such pairs as "moon-boon". Nasals were paired with voiced stops (i.e.,/m/-/b/ and /n/-/d/), to isolate the nasal feature.

c) Sustention: In 170 trials, there were 33 errors (19%). In pairing a fricative with an affricate (e.g., "shin-chin"), there were 16 errors in 40 trials (all trials were unvoiced), and 15 of these were mistaking /sh/for /ch/. This cue is difficult to hear, being almost entirely contained in the rate of onset of frication. Since the frame size was 20 msec, such transient information can be expected to be degraded. The form of the test aggravates the error performance here especially, since the listener is forced to perceive the difference between silence and a stop period at the start of an utterance.

The remaining 17 errors occurred in pairs of non-strident fricatives with stops (e.g., "vox-box"); 15 errors (of 80 trials) used voiced obstruents, and only 2 (of 50 trials) had unvoiced ones (e.g., "thick-tick"). The error rates were significantly higher for the voiced category when labial consonants were used (due to the low-amplitude noise bursts of labials). The /v/-/b/ and /ǯ/-/d/ confusions were well-known in speech perception [12], and again can be related to the slow frame rate.

d) Sibilation: In 160 trials, 18 errors (11%) were made. Most errors occurred in pairing strident fricatives with other fricatives (e.g., "sank-thank")(11 errors in 50 trials), while only 7 errors (in 110 trials) happened with pairs of affricates with stops ("juice-goose"). In the latter case, all 7 errors were mistaking /ǰ/ for /g/, and the 11 errors in the former were all mistaking /s,z/ for /th/. So perhaps the RELP algorithm does not always replicate sufficient frication energy for alveolar fricatives and affricates.

e) Graveness: 23 errors in 170 trials (14%) were found in pairing labial consonants with their corresponding dental or alveolar versions (e.g., "bid-did", "fought-thought"). The place of articulation information was well conveyed in the nasals (1 error in 20 trials), but less well in stops (9 errors in 90), and worst in fricatives (13 errors in 40). The poor fricative performance is partially due to the very difficult /f/-/θ/ discrimination. In general, these results could indicate degraded representation of the crucial formant transitions as the vowel commences after the consonant.

f) Compactness: Pairing velar consonants with non-velars (e.g., "caught-taught") yielded 8 errors in 180 trials (4%). The better performance for velars than labials reflects the stronger acoustic cues of velars (e.g., stronger noise bursts, longer voice-onset times, close second and third formants).

g) Miscellaneous: 160 trials were also included in the test pairing /r/ with /n/. Despite both consonants being sonorants with low formant frequencies, there were no errors.

In summary, the RRELP algorithm at 9.6 kb/s gave an acceptable intelligibility performance of 92.5% on initial consonant features in isolated monosyllabic words. This achievement can be compared to that of an independent 6.4 kb/s adaptive predictive coder, which attained 87% correct [13]. As a result of the above analysis of errors, two directions can be identified as ways to improve intelligibility performance: increasing the frame rate, and examining more closely the frication energy transmission for strident fricatives and affricates.

With regard to the frame rate, one must take care not to overly degrade the spectrum as the frame rate is increased. Due to the fixed transmission rate of 9.6 kb/s, it is necessary to reduce the number of bits assigned to the LP spectral representation or those given to the residual signal. Since the 4.8 kb/s RRELP showed only slight quality degradation over the 9.6 kb/s one (see below), perhaps the bits should be taken from the residual. Another possibility is to use a variable frame rate encoding scheme, to transmit spectral information only when the spectrum changes sufficiently.

### 3.3.3 Naturalness Test

The RRELP algorithm was also used to process 20 sentences (see Table 1) spoken in a quiet room (10 by an adult female, 10 by an adult male). The average duration of the sentences was about 4 seconds. Since the quality of the RRELP speech was reasonable at 4.8 kb/s, the perceptual experiment was run at the lower rate as well as at 9.6 kb/s. To judge naturalness, 9 subjects were asked to listen via earphones to a pair of sentences, and judge which was more "natural". Specifically, the instructions were "choose the one you preferred. The basis for your decision should be the quality of the speech, or its naturalness. Choose the sentence in each pair that you think sounds more natural, or has the better quality. (For instance, if voice coming over your telephone sounded like this, which would you prefer to listen to?)"

The RELP utterances were paired with log PCM utterances of the same sentence at various bit rates. Each sentence was processed by the RRELP algorithm at 4.8 and 9.6 kb/s, and also by a straightforward log PCM algorithm at rates of 24, 32, 40, and 48 kb/s (which corresponded to 3,4,5, and 6 bit log PCM at the 8000/s sampling rate). Comparing processed speech against log PCM speech is a commonly-used procedure to evaluate speech quality. Each pair of RRELP and log PCM sentences was presented once in random order to each listener. The 10 sentences by 2 speakers provided 20 sets of 8 sentence pairs each; the pairings were randomized within each set, but not across sets; the sentences spoken by the female were presented before those of the male.

The results showed the 9.6 kb/s RRELP to be as natural as 4-bit log PCM speech, and the 4.8 kb/s speech to be midway between 3 and 4 bit log PCM in naturalness. Specifically, in comparing 9.6 kb/s RRELP against 3,4,5, and 6 bit log PCM, 81%,46%,19%, and 10% of the listeners, respectively, preferred the RRELP. In judging 4.8 kb/s RRELP versus the same log PCM, 65%,26%,17%, and 4% of the subjects, respectively, thought the RRELP more natural.

There was no significant difference in the listeners' performance depending on the speaker's sex. Some RRELP sentences were judged better than others, however. In pairing 9.6 kb/s RRELP with 4 bit log PCM, 2/3 of the listeners judged the RRELP preferable on 3 sentences, while thinking the log PCM better in 8 other sentences. On the remaining 9 sentences, the subjects divided relatively evenly.

There was little inter-subject difference in performance, except for 2 subjects. One judged the 9.6 kb/s RRELP speech slightly worse than the 3 bit log PCM, whereas the other preferred the 9.6 kb/s RRELP to 6 bit log PCM, and the 4.8 kb/s RRELP to 5 bit log PCM. This illustrates the difficulty in attempting tests of naturalness and quality: if the speech is corrupted by the same process (such as adding simple noise), listeners can easily judge which samples are worse than others by listening for the noise levels or judging along a single continuous scale of degradation. But when one is asked to select preference between speech samples corrupted by different types of processing (e.g., RELP and PELP), it is hard to say which form of corruption is better. By taking 20 sentences from different speakers and using 9 subjects, the sum of the results should present an average picture of the quality of RRELP.

The BRELP algorithm as described above was implemented in software, but provided speech of lower quality than the RRELP did. For this reason and because BRELP was computationally more expensive than RRELP, the RRELP speech was exclusively used in the perceptual tests. The problem with BRELP appears to be that the approximation of the residual at the decoder, based only on the received baseband, does not represent the key perceptual aspects of e(n) as well as the low-passed version of e(n) based on the full s(n).

1) The birch canoe slid on the smooth planks.
2) Glue the sheet to the dark blue background.
3) It's easy to tell the depth of a well.
4) These days a chicken leg is a rare dish.
5) Rice is often served in round bowls.
6) The juice of lemons makes fine punch.
7) The box was thrown beside the parked truck.
8) The hogs were fed chopped corn and garbage.
9) Four hours of steady work faced us.
10) A large size in stockings is hard to sell.

TABLE 1

The 10 Sentences used in the Intelligibility Test (in order of presentation)

### 3.3.4 Algorithm Complexity

The section examines the complexities of the PELP and RRELP algorithms. The major such tradeoff is one of eliminating the pitch detector and adding the 4 filtering operations needed for RRELP. Assuming the speech to be processed is N samples long and the digital filters to have M coefficients or "taps", the requirements for the RRELP are as follows: In the analyzer, the predictor filter A and ensuing low-pass filter are not needed for PELP and require pN and MN multiply and add operations, respectively. (In these filtering procedures, multiplies and adds normally occur in pairs, and so they will be referred to henceforth simply as "operations".) In the decoder stage, the received $\hat{e}(n)$ must be low-passed, double-differenced, and high-passed. These require another 1.2MN + 2N operations, if the number of taps in the bandpass filters remains M. (The low-pass filter after the 1:5 interpolator only needs .2MN operations since .8 of the operations involve multiplies by zero.) The 2N operations for the differencer can be neglected in first-order approximations of complexity if $M >> 1$, which was the case in this study ($M = 31$). Similarly, there is a gain computation necessary in the synthesizer which takes N operations, which can also be ignored.

Also relevant is the coding and decoding of e(n) with a log PCM (or ADM) algorithm. While both PELP and RELP must code the LP spectral information, only the RELP must handle the extra e(n) or b(n). If a log PCM approach is used, the extra operations of exponentiation and logarithm are need for every sample of e(n) transmitted, in addition to the 2 multiplies normally required for linear quantization. The other operations in the RRELP algorithm are either common to PELP (i.e., the calculation, coding, and decoding of $a_i$, and the implementation of the lattice synthesizer) or are simple in comparison to the filters (e.g., the rectifier in the decoder).

These added computations must be balanced against the elimination of a pitch detector. Since there exist many pitch extraction algorithms, it is difficult to give representative figures for computation time, but, as an example, consider the SIFT algorithm [4] which is currently used in our PELP system. The SIFT procedure first low-passes and downsamples the speech signal to 1 KHz, which effectively takes 4N operations with a 4-tap digital filter. This signal is then Hamming windowed (N/4 operations, because the signal has been decimated 4:1), and an autocorrelation LP analysis performed. Since the signal is limited to 1 KHz, a 4-pole model is sufficient, which requires on the order of 5N/4 operations. A residual signal is produced via a 4-tap inverse filter (N operations), and this is also windowed (N/4 operations). The resulting signal should have major obtrusions once per pitch period (during voiced speech) corresponding to the vocal cord excitation; hence an autocorrelation is performed to find the period. This needs about 5N operations (the actual number varies directly with the range of pitch periods expected - currently, the range 2.5-15.5 msec is used (400-65 Hz)). Further operations involve peak picking and thresholding, but are computationally inexpensive. Thus, the total number of operations is about 12N, but this assumes that each section of N speech samples is used only once, and most applications would require at least a 2:1 overlap of frames for accurate pitch tracking. Thus a more reasonable estimate of operations for the SIFT algorithm would be 24N, i.e., 24 multiplies and adds per speech sample.

This compares to the 82N operations used in the current RRELP algorithm. Thus this version of the RRELP takes 3-4 times as much computation time as would a SIFT pitch detector producing pitch values at 50/s. (If one wanted pitch samples at 100/s, as many PELP algorithms do, the ratio would be 1.5-2.) The major computation for the RELP lies in the low pass filter of the analyzer and the high pass filter of the synthesizer (62N operations). A systematic study of the effects of reducing the dimensions of these filters from 31 taps has not been done, but the quality of the RELP speech does deteriorate with lower-order filters. The 31-tap filters was chosen mainly for their frequency response characteristics (described earlier). Lower order filters would have wider transition bands and less stopband attenuation (with increased aliasing degrading the speech). If computation time is a major problem, alternative filters could be used. For example, Un and Magill [3] used a 4-pole Butterworth low-pass filter.

### 3.4.0 Future Work

Informal listening tests have shown the superiority of RELP speech at 9.6 Kb/s over PELP speech but this needs to be established formally in future perceptual experiments. These tests will compare RELP and PELP speech both for undegraded input and for telephone speech.

In attempting to improve the quality of the current RELP speech without raising the transmission rate, a number of possibilities present themselves. Recently, Makhoul and Berouti [14] tried a new method of reconstructing the residual at the receiver, which uses spectral folding or translation in place of a nonlinear operation followed by spectral flattening. A major drawback of this approach is that the equal spacing of harmonics in the original residual is not preserved in the reconstruction; this adds a certain amount of ringing to the speech. Nevertheless, the basic approach has possibilities, and variations on it may yield improved speech.

In searching for ways to improve the RELP speech, most effort should be placed on better methods to transmit and reconstruct the residual. Improvements in the coding of the spectral information (i.e., the gain and the LP coefficients) will likely give only marginal improvements in RELP quality. Informal listening tests were performed using the original residual to excite a synthesis filter specified by quantized coefficients, and speech quality was only very slightly degraded using 3.2 kb/s for the spectral information. At 1.8 kb/s, there was some degradation in the form of a slight warbling during low, back vowels; at 1.2 kb/s, the spectral representation became sufficiently coarse to cause a significant amount of warbling during the vowels. The key point here is that most of the transmission rate in RELP codecs is utilized to transmit the residual, and that most of the degradation occurs in reducing the residual transmission rate to less than 9 kb/s. Since the spectral information can occupy as little as 1.5-2.0 kb/s of the total 9.6 kb/s rate for the RELP codec, it appears more fruitful to improve the residual than the spectral transmission.

The problem with the RELP speech used in the simulations is that it has a "harsh" voice quality during the vowels, sounding as if the speaker has a breathy voice. This can be directly attributed to the lack of accurate reconstruction of the residual. The higher frequencies in the synthesis filter are excited by a residual whose periodicity is accurate but which otherwise has waveform discrepancies when compared to the original.

Further research is needed to discover which aspects of the residual must be faithfully replicated at the receiver to produce perceptually-acceptable speech. It is unlikely that reconstructing the full residual from any single bandpass portion will yield any better results than were obtained using the lowpass portion, but a "sub-band" approach may have better results. The basic idea behind sub-band codecs [15] is that the frequency bands with higher amounts of energy (formants) are more perceptually important than other frequency bands; thus these bands are allocated the prime transmission bandwidth. If one tries to maintain the 800 Hz bandwidth used for the residual transmission in the simulation, an adaptive approach is necessary since the perceptually crucial first 3 formants occupy a 3000 Hz range. An adaptive approach would attempt to follow the formants, and filter the residual with a bandpass filter centered at each of the formants, transmitting perhaps a bandwidth of 300 Hz around each of the first 3 formants. At the synthesizer these 3 formants would be excited by the true residual, while the reconstructed residual would be relegated to the low spectral energy regions between formants. This approach would potentially yield better speech, at the price of requiring a crude formant extractor from the LP coefficients and doing more complex variable bandpass filtering.

Alternatively one could try waveform coding of the residual [16-18], attempting to preserve the perceptually important aspects of the residual. Standard waveform techniques such as ADPCM, CVSD and ADM are not likely to be useful at rates under 9.6 kb/s in preserving the waveform accurately, but it may well be that certain distortions in the residual waveform do not yield much perceptual difference in the resulting synthesized speech. Informal listening tests using various simplifications of the residual illustrated some possibilities. For example, a linearization of the residual between waveform peaks caused little perceptual change.

Since the basic problem lies in the voiced regions, the effort for better residual encoding must go there. The LP model for voiced speech suggests that the residual should approximate a train of impulses spaced at the fundamental frequency, and indeed actual residual signals are well modeled by periods of large waveform excursions follwed by small irregular oscillations. Preserving the beginning of each period while replacing the rest with a zero waveform yielded clear, but raspy speech. Such an approach could reduce the residual transmission rate to below 9 kb/s. Initial trials produced a crisper sounding speech than the original simulation, eliminating the harsh, breathy quality; however, the raspiness presents a problem on about the same perceptual level as the previous harshness.

Since the residual waveform is more accurately preserved in this new method, the final speech waveform is a more accurate replica of the original than with the earlier RELP approach. This provides encouragement that some form of waveform coding on the residual signal may yet prove a better approach to residual reconstruction, and hence better quality RELP speech.

## 4. SUB-BAND CODING OF SPEECH

### 4.1.0 Theory

For the digital transmission of a speech signal, the signal must be sampled and each sample quantized. If each sample is independently quantized as in conventional pulse code modulation (PCM), the number of quantizer levels must be large (usually 256) for high quality speech reproduction. The technique considered here, sub-band coding, allows the use of a much smaller number of quantizer levels without impairing speech quality. This smaller number of quantizer levels allows for a large reduction in the transmission rate for digitally coded speech.

The advantages of sub-band coding can be explained as follows. As long as the number of quantization levels is large, the noise due to quantization is nearly white and is essentially uncorrelated with the input signal. As the number of quantization levels is reduced, these properties no longer hold. The quantization noise is annoyingly correlated with the input signal. Sub-band coding tries to alleviate this problem by partitioning the input spectrum into sub-bands. Each sub-band is then independently quantized (see Fig. 3). In this approach, the quantization noise in a sub-band is still correlated with the signal in that sub-band but to a lesser degree with the signals in other sub-bands. The overall effect when the sub-bands are combined is to reproduce speech in which the perceptual effects of quantization noise are very much reduced as compared to conventional full band coding of speech. There is additional flexibility in sub-band coding in that the number of quantization levels can be different for different sub-bands.

While related to coding schemes which have long histories, the form of sub-band coding considered here was first considered by Crochiere, Webber and Flanagan [15].

Several forms of sub-band coding have been evaluated at low bit rates (16 kb/s and 9.6 kb/s). The ultimate purpose of this work is to determine the practicality of this technique for high quality transmission of speech at low bit rates. The sub-band coders were simulated on a digital computer. While this approach does not approach real time operation, the flexibility of the simulation allows for system parameters to be modified relatively easily. The resultant speech was then subjectively compared to that obtained by using conventional PCM.

### 4.1.1 Choice of Sub-Bands

Consider the two extremes. The first would be one sub-band. This is full band coding as practiced in PCM or differential PCM. The transmission rate may be reduced by taking advantage of the correlation between time samples. However, as mentioned earlier, the problem of signal correlated non-white quantization noise manifests itself as the number of quantization levels is reduced.

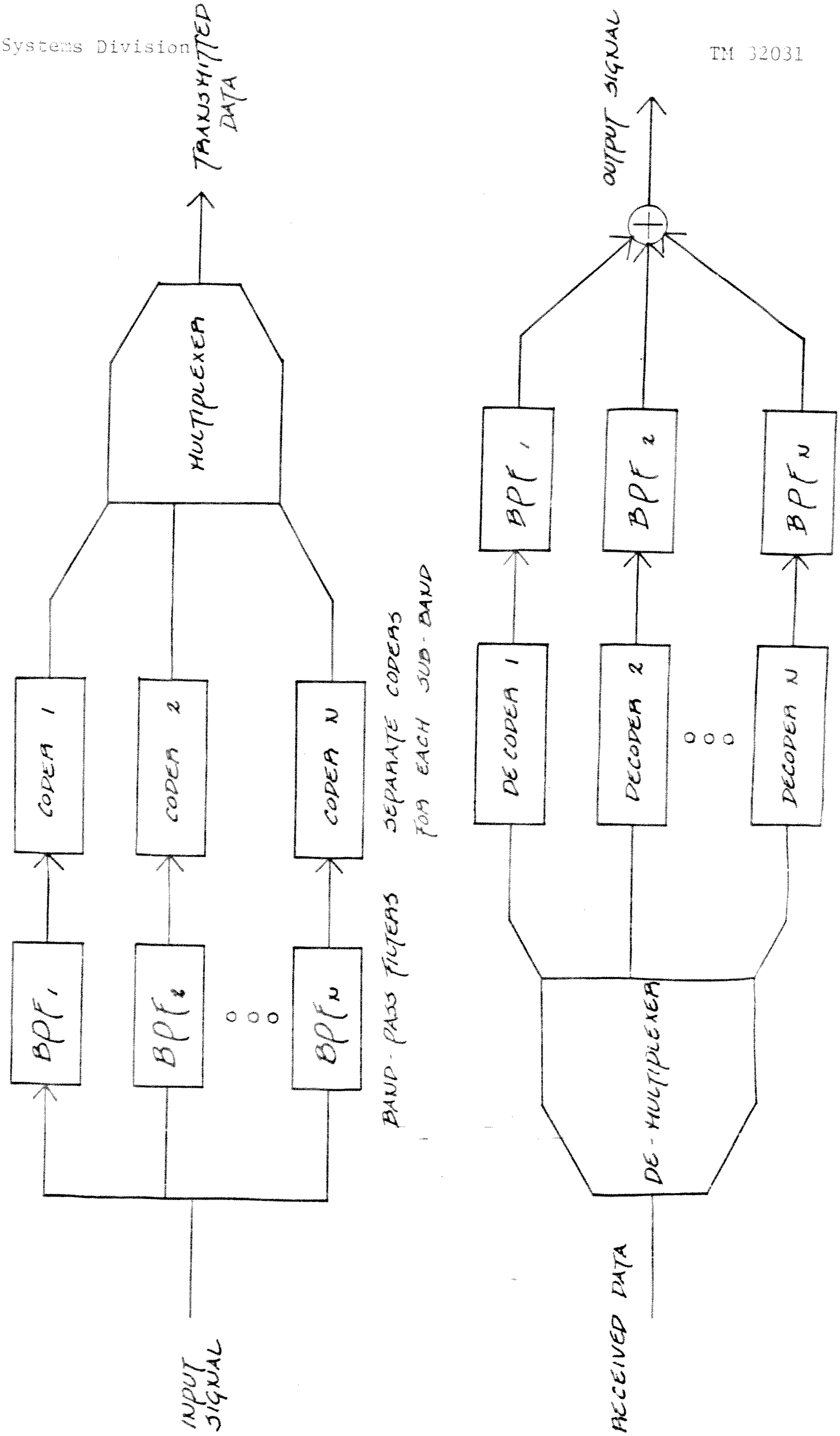


Figure 3 - Block Diagram of a Sub Band Coder

The other extreme is to employ a large number of narrow sub-bands. The output of each sub-band filter represents the response at the frequency corresponding to its centre frequency. This is then a form of frequency domain coding. The quantization of these frequency domain samples does lead to quantization noise which is less offensive perceptually than full band coding. However if a low overall bit rate is to be achieved, the output of each sub-band must be sampled at a low rate. Thus the time correlation between samples from a sub-band filter is very low.

The compromise employed in sub-band coding is to use four or five sub-bands of varying widths. With this number of sub-bands, adaptive quantizers which adapt to the average short time energy in each sub-band are practical since the sampling rate for each sub-band is still reasonably high. The width of each sub-band is chosen relative to its perceptual importance in reproducing speech. For instance, the Articulation Index (AI) for each sub-band could be chosen to be the same [19]. These perceptual considerations point to narrower sub-bands at lower frequencies. Sampling considerations will also affect the exact choice of sub-bands.

### 4.1.2 Sampling of Sub-Band Signals

Consider the problem of efficiently sampling the band-pass signal corresponding to a sub-band of the original signal.

Let the band-pass signal be bandlimited to frequencies between $f_U$ and $f_L$ (see Fig. 4). Direct sampling of this signal causes the frequency response to be repeated at the sampling rate. If no overlap of the repetitions of the basic frequency response is to occur (no aliasing), the sampling rate should satisfy

~~$f_S = 2W \; (((f_U - f_L) / W) / \lfloor (f_U - f_L) / W \rfloor )$~~

$$f_s = 2W \frac{\frac{f_u}{W}}{\left\lfloor \frac{f_u}{W} \right\rfloor}$$ where $W = f_u -$

, where $\lfloor a \rfloor$ denotes the greatest integer not bigger than a.

If the band-pass signal extends exactly from $nf_s/2$ to $(n+1)f_s/2$ for some integer n, the signal can be sampled at a rate $2(f_U-f_L)$. The additional constraint that the sampling rate for each sub-band be a sub-multiple of a higher rate will be imposed later.

### 4.1.3. Frequency Translation of Sub-Bands

If the frequency range of a sub-band does not fall between multiples of $f_s/2$, the sub-band can be frequency translated before sampling. By appropriately shifting the sub-band, the sub-band can always be sampled efficiently at the rate $2(f_U-f_L)$, where $f_U-f_L$ is the bandwidth of the sub-band.

The most general frequency translating arrangements are shown in Fig. 5. The first uses a Hilbert Transform; the second uses complex demodulation. The equivalence of these circuits is discussed in Appendix A. If the shift frequency $f_c$ is such that

$$|f_c| \geq f_U / 2 , \tag{1}$$

the simpler circuit shown in Fig. 6 can be used. In this case, only a band-pass filter is required.

Generally, shifting the sub-bands to between 0 and $f_U-f_L$ is possible with the simplified arrangement of Fig. 6 for all but the lowest sub-band. For these cases the filter in Fig. 6 is a low-pass filter. For the lowest sub-band, the requirement given by (1)

$$|f_c| \geq f_U / 2$$

may not be met. It is then possible to shift the sub-band to between $n(f_U-f_L)$ and $(n+1)(f_U-f_L)$ for some value of n other than zero. If n=-1 (or more generally negative), the frequency components in the sampled signal will be inverted but (1) will always be satisfied.

4.1.4 Sampling

Until now, the formulation of the sub-band coder has proceeded as if the filtering and frequency translation were all analogue processes to be followed by a sampler. In practice it is preferable to sample the signal before filtering. Thus the filtering and modulation become digital operations.

The sampling of the output of the frequency translator then becomes an operation to change the sampling rate. While in principle it is possible to change the sampling rate by any rational factor, sub-sampling is far easier and generally does not involve any substantial compromises in the choice of sub-bands. Thus the basic scheme, so far, is to digitally filter the signal into sub-bands and then sub-sample the signal at the rate appropriate to each sub-band.

4.1.5 Quantizers

Since the samples for each sub-band are from a sub-sampled sequence, the sample-to-sample correlation is low. Thus no benefit is to be gained by differential encoding schemes. Another approach is to use a quantizer with a variable step size [20]. The step size of the quantizer is varied according to the rule

$$\Delta_n = \Delta_{n-1} M_q$$

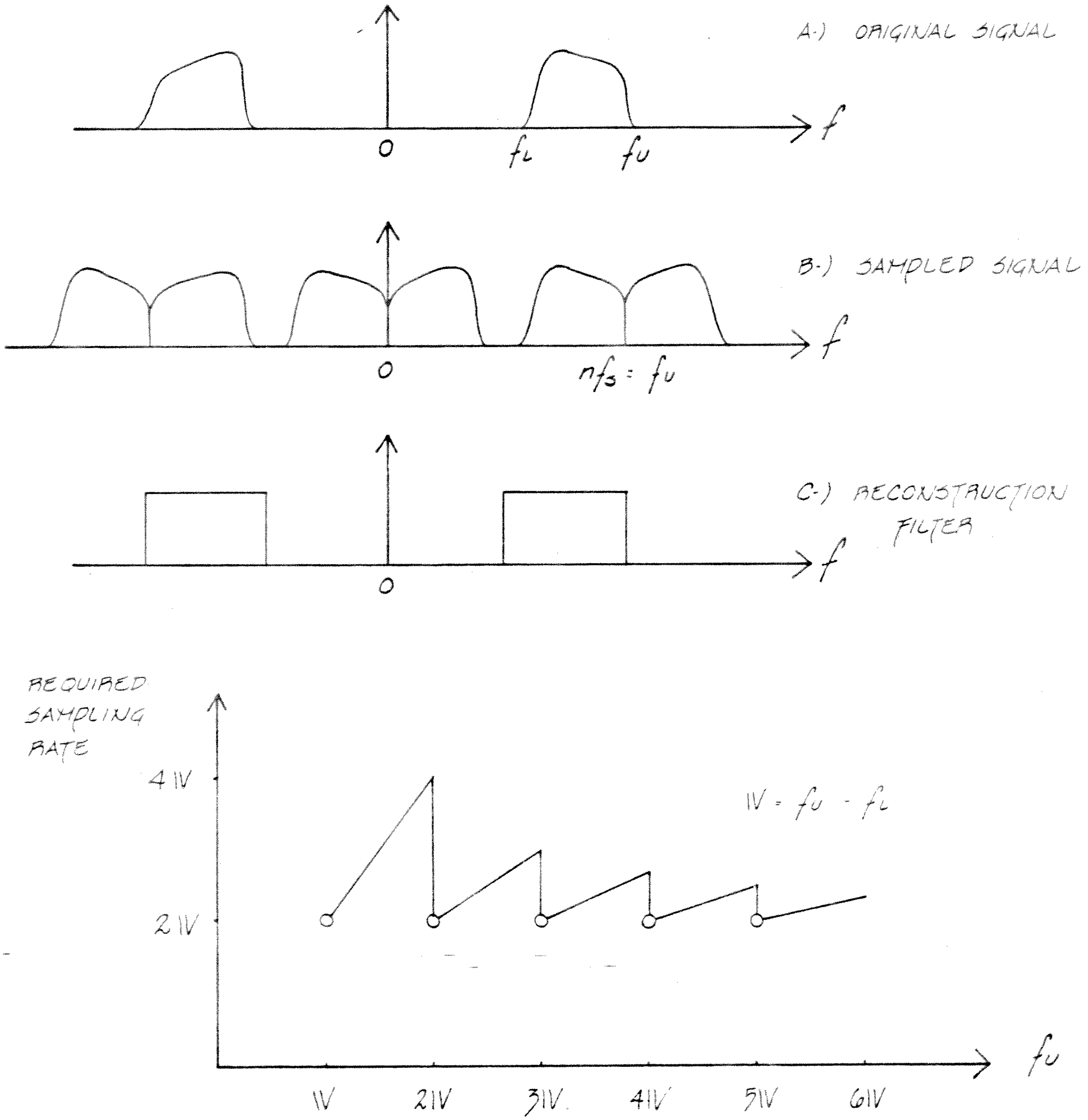


Figure 4 - Sampling Band-Pass Signals

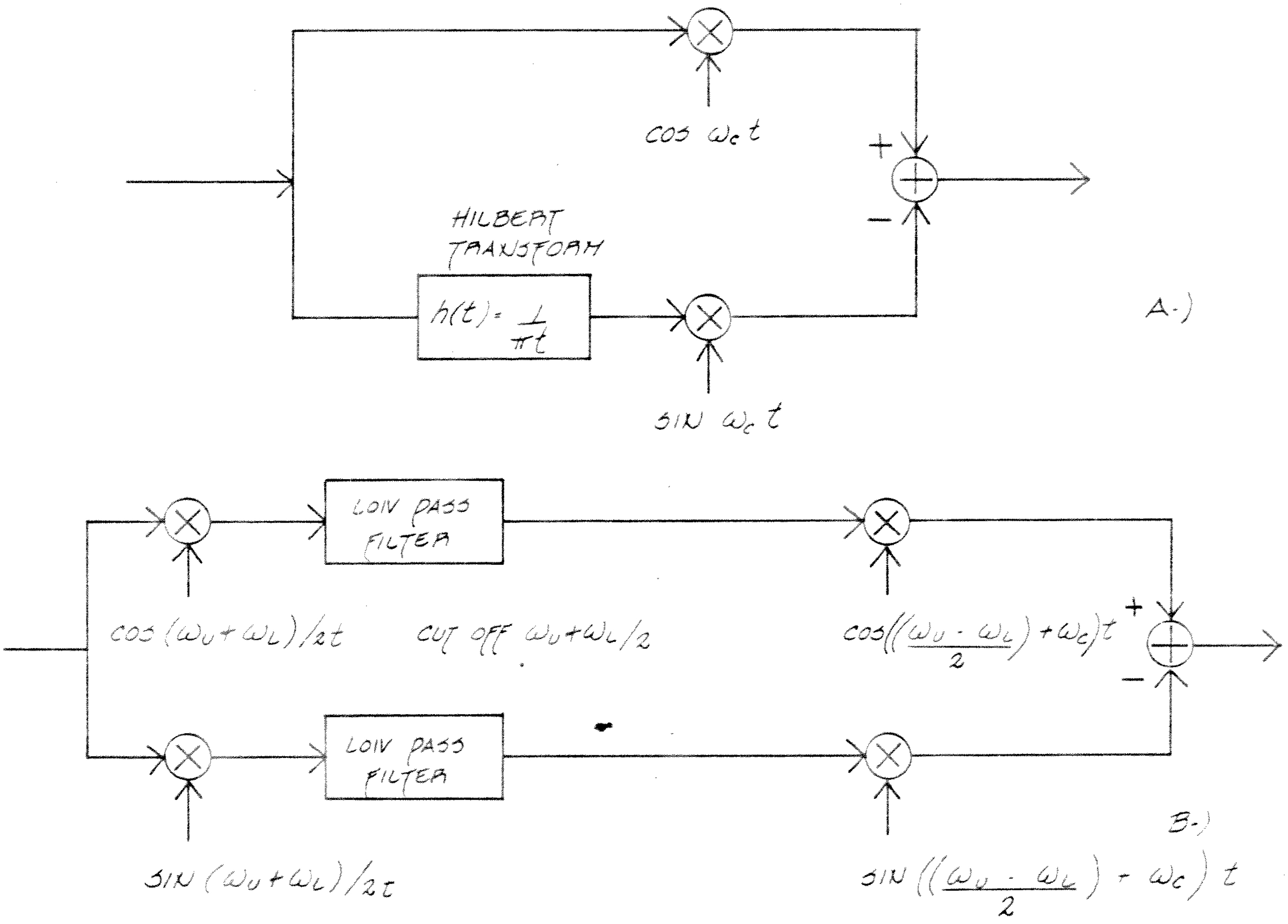

$\cos \omega_c t$
+
−
HILBERT TRANSFORM
$h(t) = \frac{1}{\pi t}$
A-)
$\sin \omega_c t$
LOW PASS FILTER
$\cos(\omega_U + \omega_L)/2t$
CUT OFF $\omega_U + \omega_L/2$
$\cos((\frac{\omega_U - \omega_L}{2}) + \omega_c)t$
+
−
LOW PASS FILTER
B-)
$\sin(\omega_U + \omega_L)/2t$
$\sin((\frac{\omega_U - \omega_L}{2}) + \omega_c)t$

Figure 5 - Frequency Translation Circuits

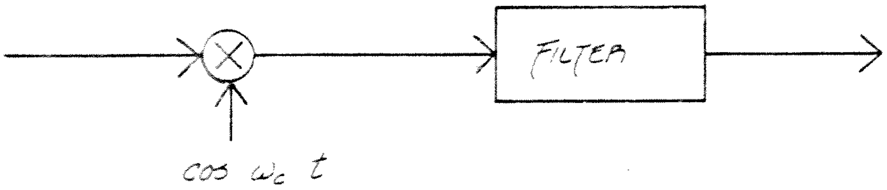


Figure 6 - Simplified Frequency Translation Circuit

Where $M_q$ is a multiplication factor which depends on the quantizer level used to quantize the previous sample. $M_q$ is larger than one for the outer levels of a quantizer and less than one for the inner levels of the quantizer. Table 2 gives appropriate values for $M_q$. These values were found to be a good compromise for several speakers and sentences.

The ratio of the largest step size to the smallest step size was kept at 100 to 1. The reason for limiting the minimum step size is to reduce the time required to recover from a long quiet period. This ratio was found adequate for the sentences used. The absolute range of step sizes for a given sub-band was determined experimentally but was found to correspond roughly with the energy in each sub-band.

### 4.2.0 Simulations

The specific choice of sub-bands is influenced by several considerations. The narrowest sub-band will have little sample-to-sample correlation. The adaptive quantizers used control their step-sizes in response to the input signal level. The minimum sampling rate (that of the narrowest sub-band) should be such that the quantizer can follow trends in the average signal level for that sub-band. In addition the widths of the sub-bands can at best be of the order of the widths of the transition regions of the filters used. For practical filters the transition widths are in the order of 100 Hz. Thus the narrowest sub-band width should be greater than 100 Hz.

Two filtering schemes were used. The first used four sub-bands; the second used five sub-bands.

### 4.2.1 Four Sub-Bands

The four sub-bands were chosen to cover the frequency range from 200-3200 Hz in four bands. Each sub-band was chosen to contribute about 20% to the Articulation Index [15]. The original speech was sampled at 10 kHz. Sub-bands were sub-sampled from this rate. Fig. 7 shows the sub-bands used and Table 3 gives the sub-band parameters. The overall processing for a single sub-band is shown in Fig. 8. The filters were implemented as finite impulse response (FIR), linear phase, filters. Linear phase is appropriate since the responses from individual sub-bands will be summed - severe phase distortion at the sub-band edges is undesirable. For the top 3 sub-bands, the range of frequencies was translated down to near zero frequency. In this case FILTER 1 is a high-pass filter and FILTER 2 is a low-pass filter (see Fig. 8). Note that reconstructing a sub-band consists of merely reversing the procedure. That is, the quantized samples are interpolated by first increasing the sampling rate and then filtering. This is followed by frequency translation and more filtering.

| no. levels | multipliers Mq |
|---|---|
| 4 | 0.845, 1.96 |
| 8 | 0.845, 1.0,1.0,1.4 |

Table 2 - Quantizer Multipliers

The reconstruction filters can be the same filters as used to generate the samples of the sub-band. The lowest sub-band was translated up in frequency before sampling. For this sub-band, FILTER 1 is a low-pass filter and FILTER 2 is a high-pass filter. The whole system gives virtually distortionless (though bandlimited) output when the quantizers are removed from the system.

As shown in Table 3, the number of quantizer levels was chosen to give an overall transmission rate near 16 kb/s. A number of sentences spoken by both male and female speakers were processed using the sub-band coder. Table 4 lists the test material. The measured signal-to-noise ratio for this material was between 11 and 13 db. Informal listening tests were performed to compare this coder with mu-law companded PCM (mu=100). The PCM coder was presented with the same speech material sampled at an 8 kHz rate. At a rate near 16 kb/s, the sub-band coder was superior to 4 bit mu-PCM (32 kb/s, SNR=15dB), equally preferable to 5 bit mu-PCM (40 kb/s, SNR=21 dB) and somewhat inferior to 6 bit mu-PCM (48 kb/s, SNR=27dB). The subjective degradations in 16 kb/s sub-band coding are

1) bandlimiting the output speech spectrum (200-3200 Hz bandwidth)

2) a background hiss

3) a slight gargling effect superimposed on speech

The fact that the quantizing noise manifests itself primarily as a background hiss, indicates that the sub-band coding scheme is achieving the desired effect which is to make the quantizing noise perceptually uncorrelated with the input signal. This effect is very noticeable when sub-band coding is compared to mu-PCM. The mu-PCM has an annoying quantization-noise/signal interaction at low bit rates. This also accounts for the sub-band coder being prefered even when its SNR is worse than that for mu-PCM.

### 4.2.2 Five Sub-Bands

An alternate sub-band coder using 5 sub-bands was also evaluated. In this coder, the sub-bands were chosen to have their band edges at multiples of half the sampling rate. Thus direct sampling of the sub-bands was possible, obviating the need for frequency translation. The computational load of the filtering is reduced since the filter output need only be calculated at the sub-sampled rate (see Fig. 9). The filter order was halved compared to the previous design by using a 255 tap FIR band-pass filter instead of separate low-pass—high-pass filters. These factors pointed to an easier implementation but slightly poorer filter responses.

| SUB-BAND | FREQUENCY RANGE Hz | TRANSLATION FREQUENCY Hz | SAMPLING RATE Hz | BITS | QUANTIZER RELATIVE STEP SIZE |
|---|---|---|---|---|---|
| 1 | 137-762 | 1113 | 1250 | 3 | 1.0 |
| 2 | 648-1362 | 648 | 1428 | 3 | 0.8 |
| 3 | 1248-2081 | 1248 | 1667 | 2 | 0.25 |
| 4 | 1985-3235 | 1985 | 2500 | 2 | 0.2 |

TOTAL RATE = 16.4 kb/s

Table 3 - Four Band Coder

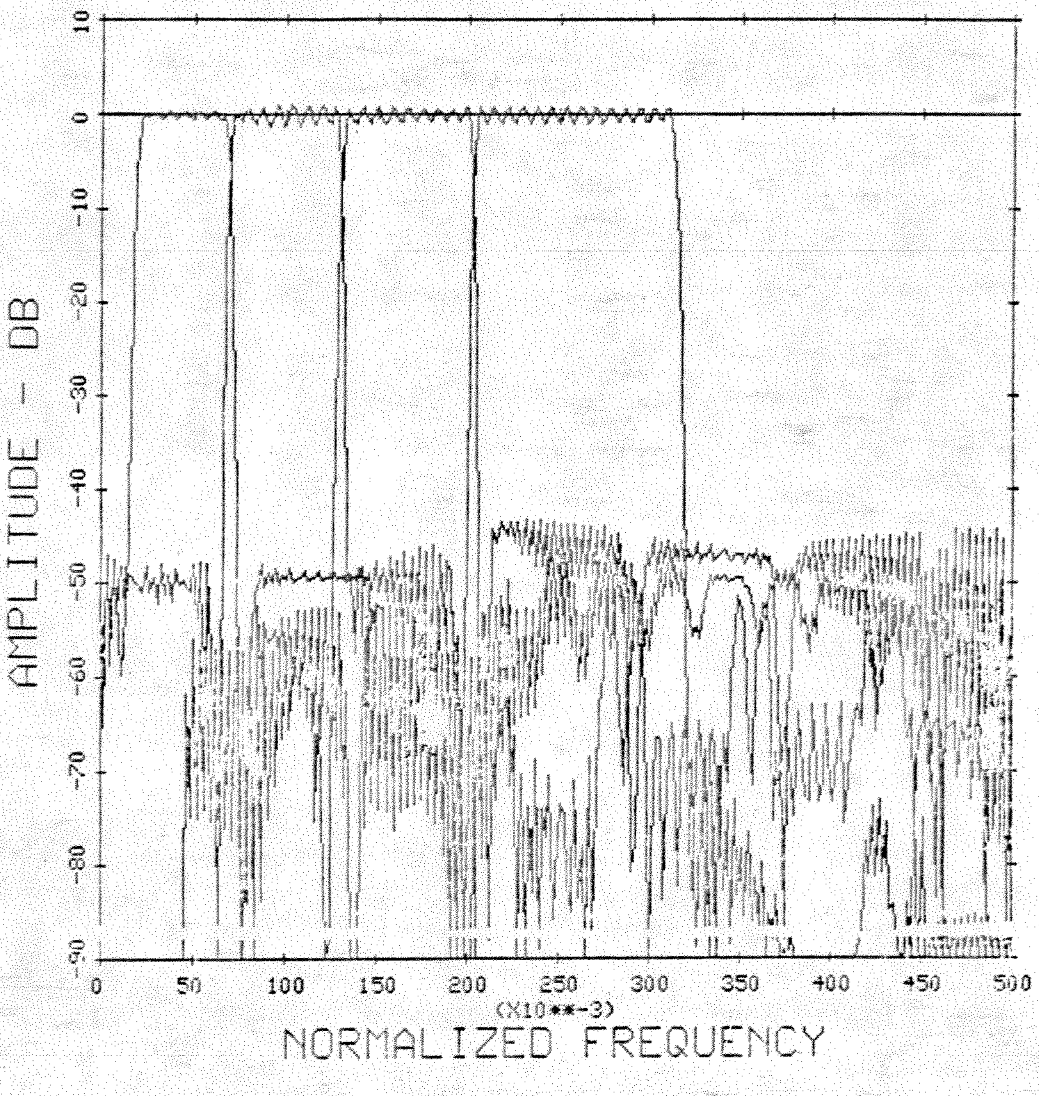


Figure 7 - Four Band Coder Frequency Response

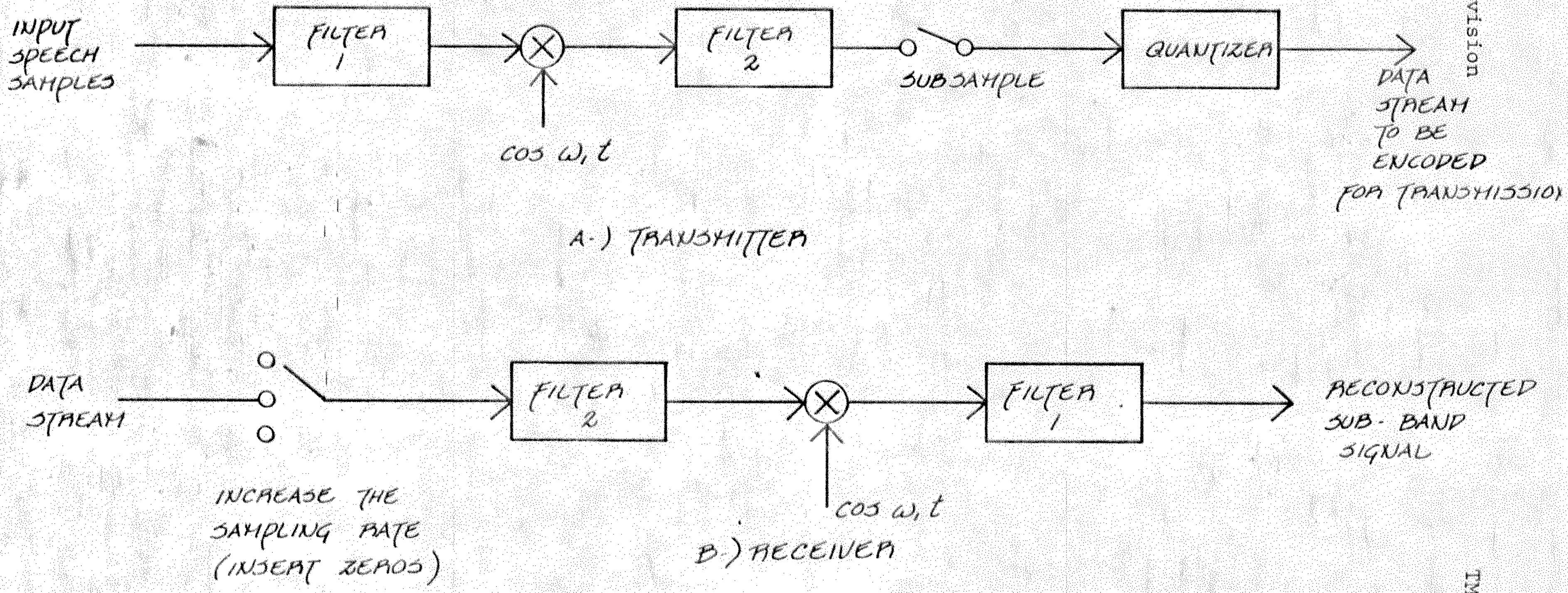


Figure 8 - Sub-Band Coder Using Frequency Translation

1. The pipe began to rust while new.
2. Thieves who rob friends deserve jail.
3. Oak is strong and also gives shade.
4. Cats and dogs each hate the other.

Each sentence was spoken by a male and a female speaker.

Table 4 - Test Sentences for Sub-band Coding

Due to the restriction that sub-bands lie between multiples of half the sampling rate and that the sampling rates themselves be sub-multiples of the original sampling rate, it was not possible to overlap the higher sub-bands. The overall response shows small dips at the transitions between sub-bands (see Fig. 10). In order to place the sub-bands at a more advantageous location with respect to the speech spectrum, the sampling rate of the original speech was increased to 10.67 kHz (2/3 X 16 kHz). Thus the input speech material was recorded at 16 kHz, interpolated by a factor of 2, subsampled by a factor of 3 before being applied to the sub-band coder.

For this sub-band coder, at a rate near 16 kb/s, the SNR was measured to be around 13 dB, an average increase of a marginal 1 dB over the 4 band coder. Subjectively, the quality of both coders was very similar.

This form of the coder is of interest since it demonstrates the simplifications possible. The greatest computational load in sub-band coding is the filtering operation. If the filters are implemented using charge-coupled-device (CCD) technology, the remaining computations are easily handled. CCD transversal filters are now commercially available in 64 tap formats. It is expected that in the near future CCD transversal filters could be employed in practical sub-band coders.

### 4.3.0 Coding at 9.6 Kb/s

To reach lower bit rates, either the bandwidth of the system or the number of bits available to a sub-band must be decreased.

Simply lowering the number of bits allocated to each sub-band results in output quality which degrades rapidly as the number of bits is reduced. Crochiere et al [15] opted for also reducing the bandwidth of the system by introducing gaps between the sub-bands. For instance at 9.6 kb/s, their sub-band coder had a 100 Hz gap near 1 kHz and a 320 Hz gap near 1700 Hz. The justification given is that there is a trade-off between the echo effect thus introduced and the quantization noise that would otherwise be introduced by the overly coarse quantizers. The use of spectral gaps has been avoided here because a system employing gaps cannot even crudely reproduce tones in those frequency ranges.

Preliminary work on sub-band coding at 9.6 kb/s has used the 4 sub-band arrangement of Fig. 7 with bit assignments of 2,2,1,1 for the sub-bands for an overall rate just below 9.6 kb/s. The 1-bit coders for the upper sub-bands employed CVSD (continuously variable slope delta modulation). In CVSD the step size of a 1-bit differential quantizer is increased if slope overload is probable (as indicated by 3 output terms of the same sign). If slope overload is not indicated the step size decays to a minimum value.

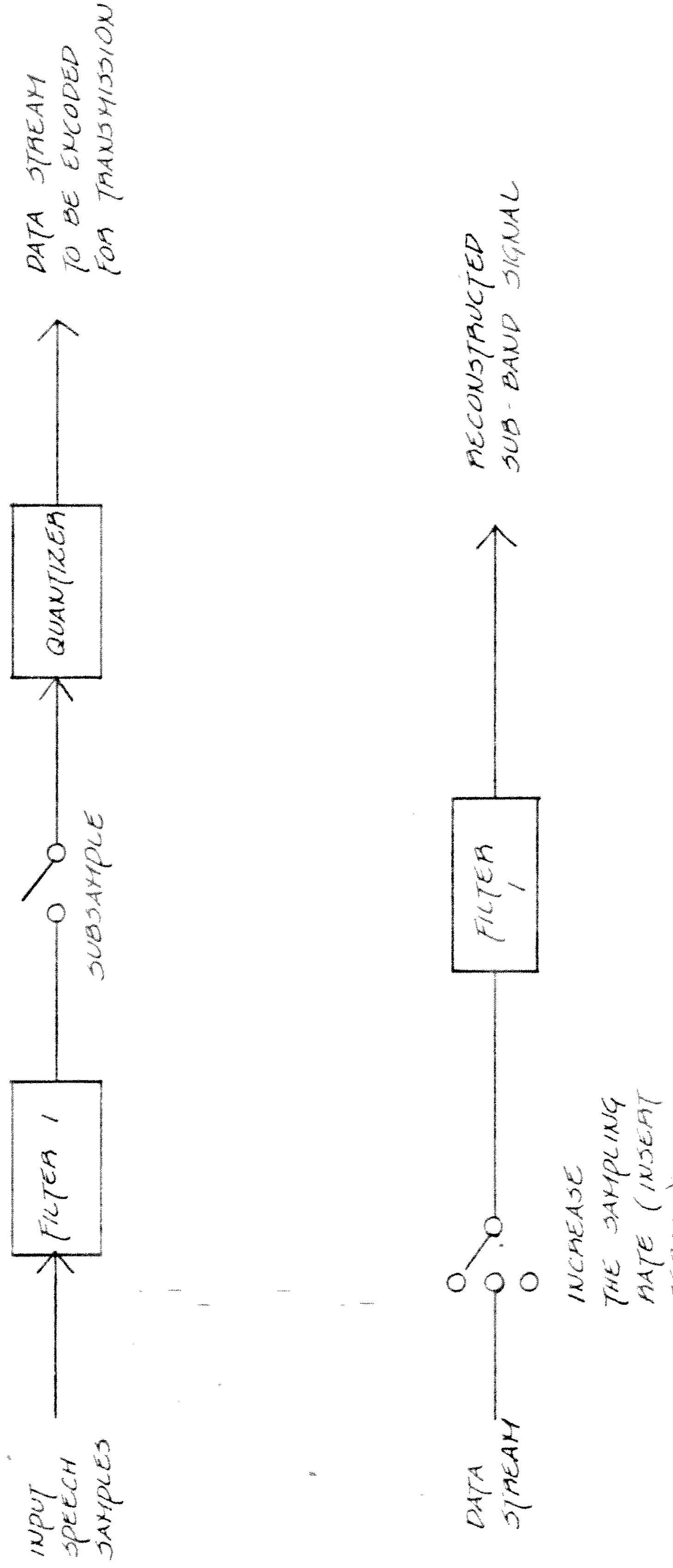

INPUT SPEECH SAMPLES
FILTER 1
SUBSAMPLE
QUANTIZER
DATA STREAM TO BE ENCODED FOR TRANSMISSION
DATA STREAM
INCREASE THE SAMPLING RATE (INSERT ZEROS)
FILTER 1
RECONSTRUCTED SUB-BAND SIGNAL

Figure 9 - Sub-Band Coder Using Direct Sampling

| SUB-BAND | FREQ. RANGE<br>Hz | SAMPLING RATE<br>Hz | QUANTIZER<br>BITS | RELATIVE<br>STEP SIZE |
|---|---|---|---|---|
| 1 | 178-356 | 356 | 4 | 0.6 |
| 2 | 296-593 | 593 | 4 | 1.0 |
| 3 | 533-1067 | 1067 | 3 | 0.25 |
| 4 | 1067-2133 | 2133 | 2 | 0.07 |
| 5 | 2133-3200 | 2133 | 2 | 0.02 |

Total Rate = 15.5 kb/s

Table 5 - Five Band Coder

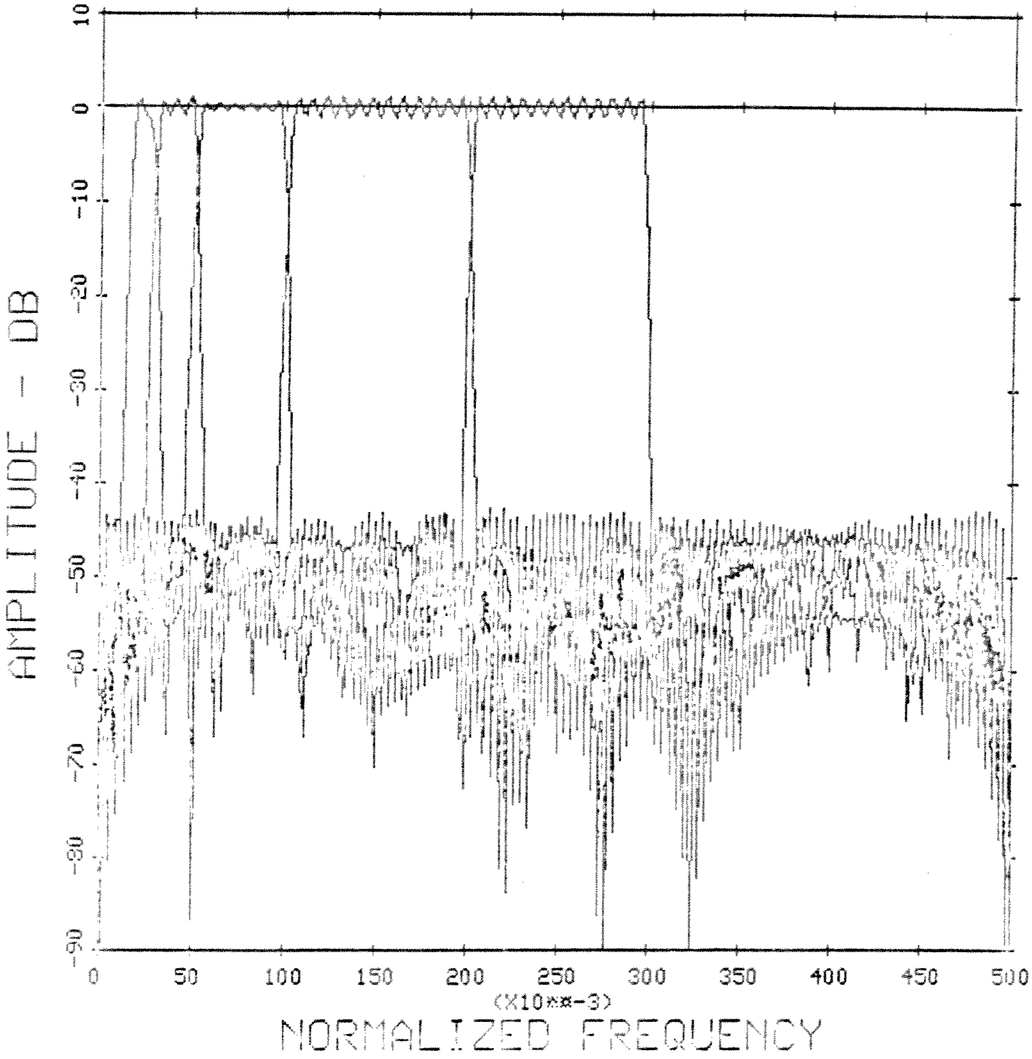


Figure 10 - Five Band Coder Frequency Response

The coarser quantizers lead to degraded performance even though the speech is very intelligible. The measured SNR is 6 to 7 dB. The subjective impairments are that the quantization noise is now somewhat signal correlated and that the "gargling" effects are amplified. The overall quality was judged to be slightly inferior to 32 kb/s mu-PCM. It is expected that the quality of sub-band coding at 9.6 kb/s can be further improved.

### 4.4.0 Future Work

The work to date has produced a good quality robust coder at rates near 16 kb/s. There is a threshold effect that causes a rapid deterioration of quality as the rate is pushed below this level. Several techniques are being evaluated which should help produce better quality speech at 9.6 kb/s. They are as follows:

1) Improved 1 bit quantizers. The adaptive quantizer strategy given earlier fails for 1-bit quantizers since there is no inherent over-range or under-range information in a single 1 bit sample. Yet 1 bit quantizers are necessary for some of the low energy sub-bands. The dynamic action of these quantizers should be helped by incorporating over-range or under-range information from adjacent sub-bands.

2) Some of the deleterious effects of leaving gaps between sub-bands may be lessened by allowing the energy in the gaps to alias into the sub-bands. In this way tones in the gaps will be reproduced albeit at a slightly different frequency.

## Appendix A - Frequency Translation

Any real signal a(t) can be represented as follows. First form its pre-envelope.

$$a^{+}(t) = a(t) + j\,\hat{a}(t)\ ,$$

where $\hat{a}(t)$ is the Hilbert transform of a(t).

Define the complex envelope of a(t) as $a^{o}(t) = a^{+}(t)\ e^{-jw_c t}$

then $a(t) = \mathrm{Re}\ (a^{o}(t)\ e^{\ jw_c t}\ )$.

This is a representation of an arbitrary real signal in terms of a "centre" frequency $f_c$ and a complex envelope $a^{\circ}(t)$. If a(t) is bandpass, $a^{\circ}(t)$ is low pass. Fig. A-1 illustrates the frequency relationships between a(t) and its pre-envelope and complex envelope.

The problem of translating frequencies can be posed as finding

$$b(t) = \mathrm{Re}\ (a^{o}(t)\ e^{jw_o t}\ )$$

given a(t). Then b(t) is a frequency translated version of a(t).

The signal b(t) can be rewritten as

$$b(t) = a(t)\ \cos(w_o - w_c)t - \hat{a}(t)\ \sin(w_o - w_c)t. \qquad (A-2)$$

This immediately suggests the circuit shown in Fig. A-2. This circuit can be rearranged by using post-filtering instead of pre-filtering. Consider the two circuits in Fig. A-3. These are equivalent if the following equalities are satisfied.

$$h_1(t) = g_1(t)\cos w_1 t + g_2(t)\sin w_1 t \qquad h_2(t) = g_2(t)\cos w_1 t - g_1(t)\sin w_1 t$$

or

$$g_1(t) = h_1(t)\cos w_1 t - h_2(t)\sin w_1 t \qquad g_2(t) = h_1(t)\sin w_1 t + h_2(t)\cos w_1 t$$

Using these equivalences the circuit shown in Fig. A-2 can be transformed to that shown in Fig. A-4. The lower branch of this circuit can be omitted if

$$|f_c - f_U| \geq f_U / 2$$

where $f_U$ is the highest frequency component in a(t).

Another possibility for deriving the signal b(t) from a(t) is suggested by (A-1).

$$b(t) = \mathrm{Re}\ (a^{o}(t))\ \cos w_o t - \mathrm{Im}\ (a^{o}(t))\ \sin w_o t$$

The signals $\mathrm{Re}\ (a^{o}(t))$ and $\mathrm{Im}\ (a^{o}(t))$ can be derived from a(t) by using complex demodulation as shown in Fig. A-5. In this implementation the Hilbert Transform filter of Fig. A-2 is replaced by a pair of modulators and a pair of low-pass filters.

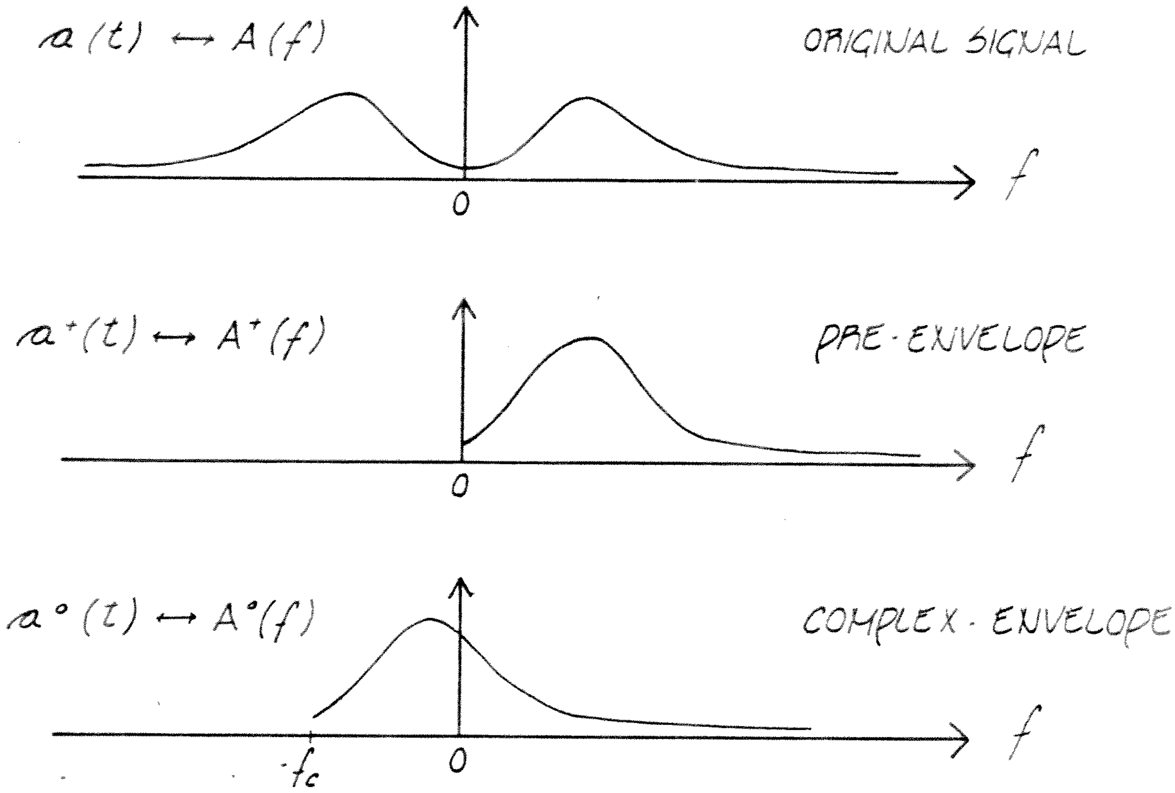


Figure A-1 - Pre-Envelope and Complex Envelope

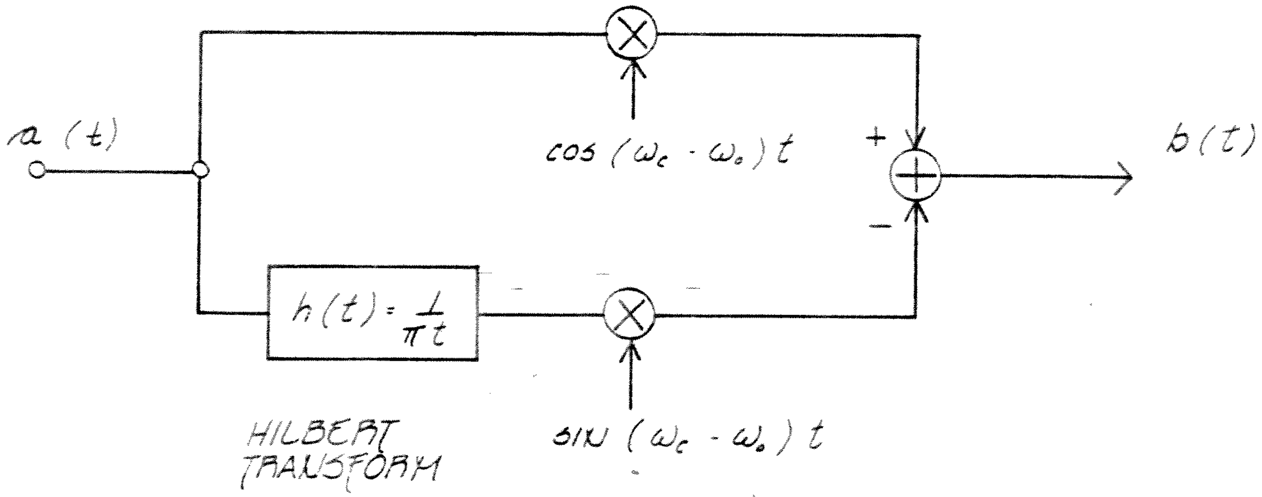
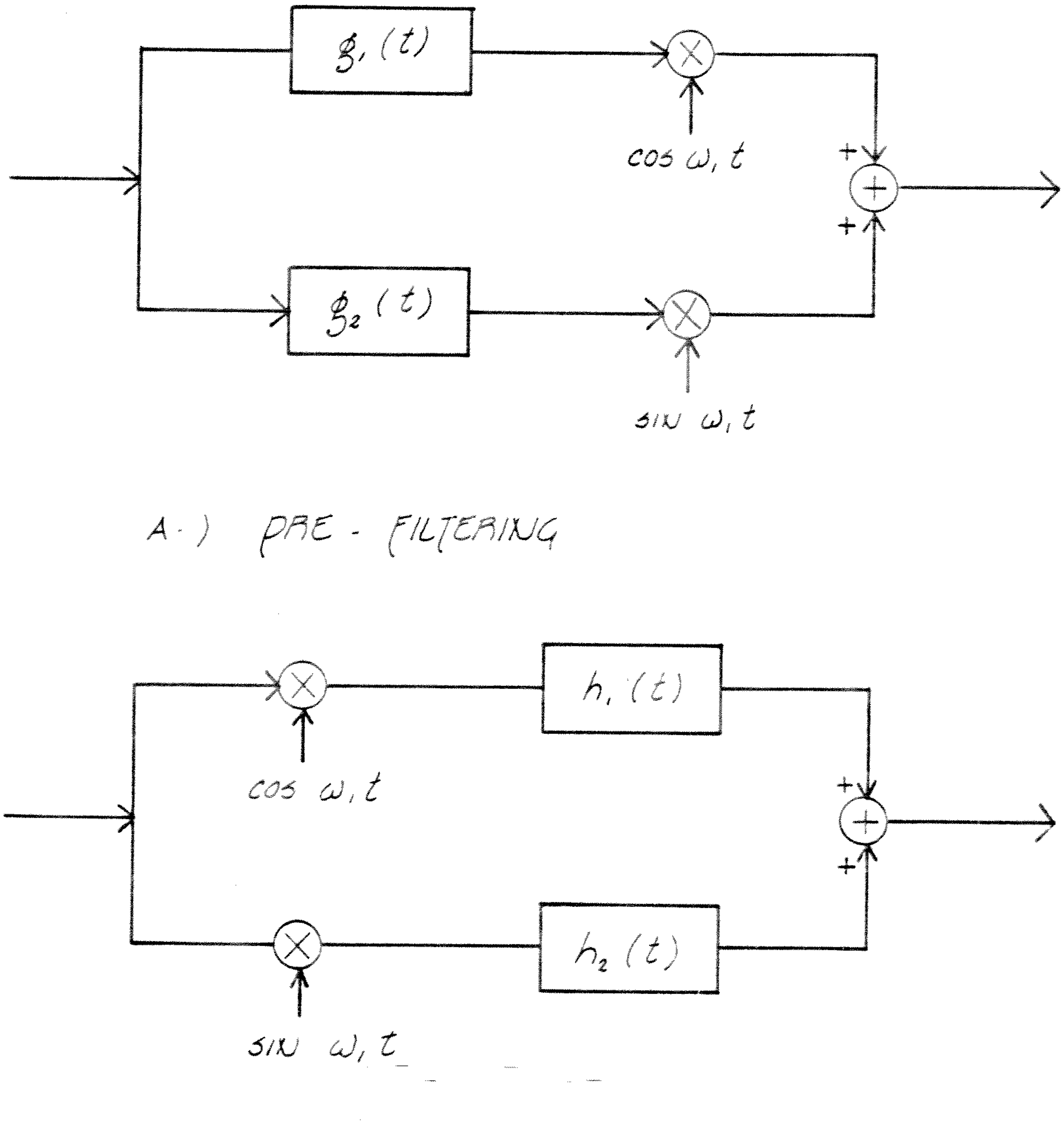


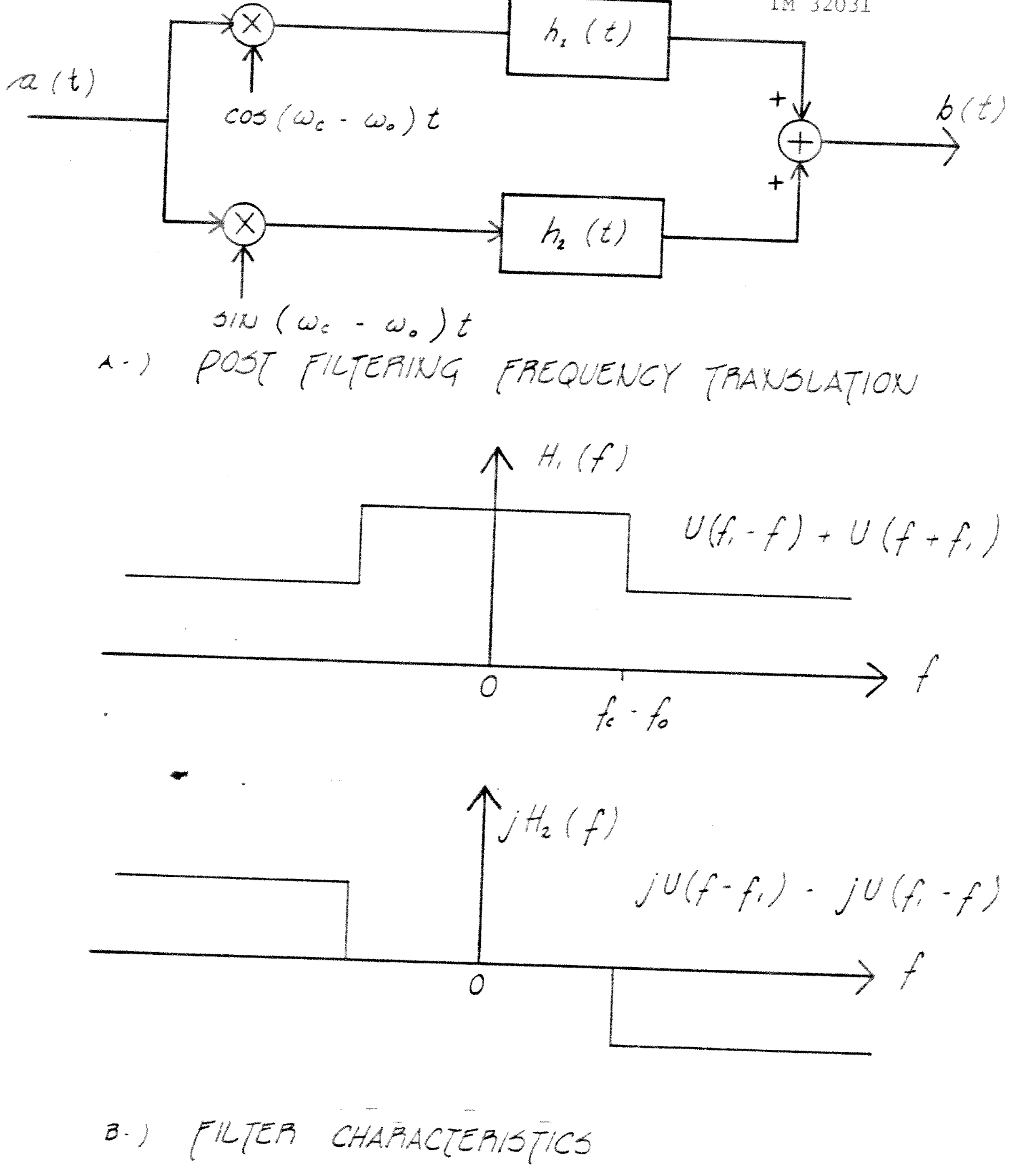
Figure A-2 - Frequency Translation Circuit



A.) PRE - FILTERING

B.) POST- FILTERING

Figure A-3 - Equivalence between Pre-Filtering and Post-Filtering


$a(t)$
$\cos(\omega_c - \omega_o)t$
$h_1(t)$
+
+
$b(t)$
$h_2(t)$
$\sin(\omega_c - \omega_o)t$
$H_1(f)$
$U(f_1 - f) + U(f + f_1)$
0
$f$
$f_c - f_o$
$jH_2(f)$
$jU(f - f_1) - jU(f_1 - f)$
0
$f$

A-) POST FILTERING FREQUENCY TRANSLATION

B-) FILTER CHARACTERISTICS

Figure A-4 - Frequency Translator Using Post-Filtering

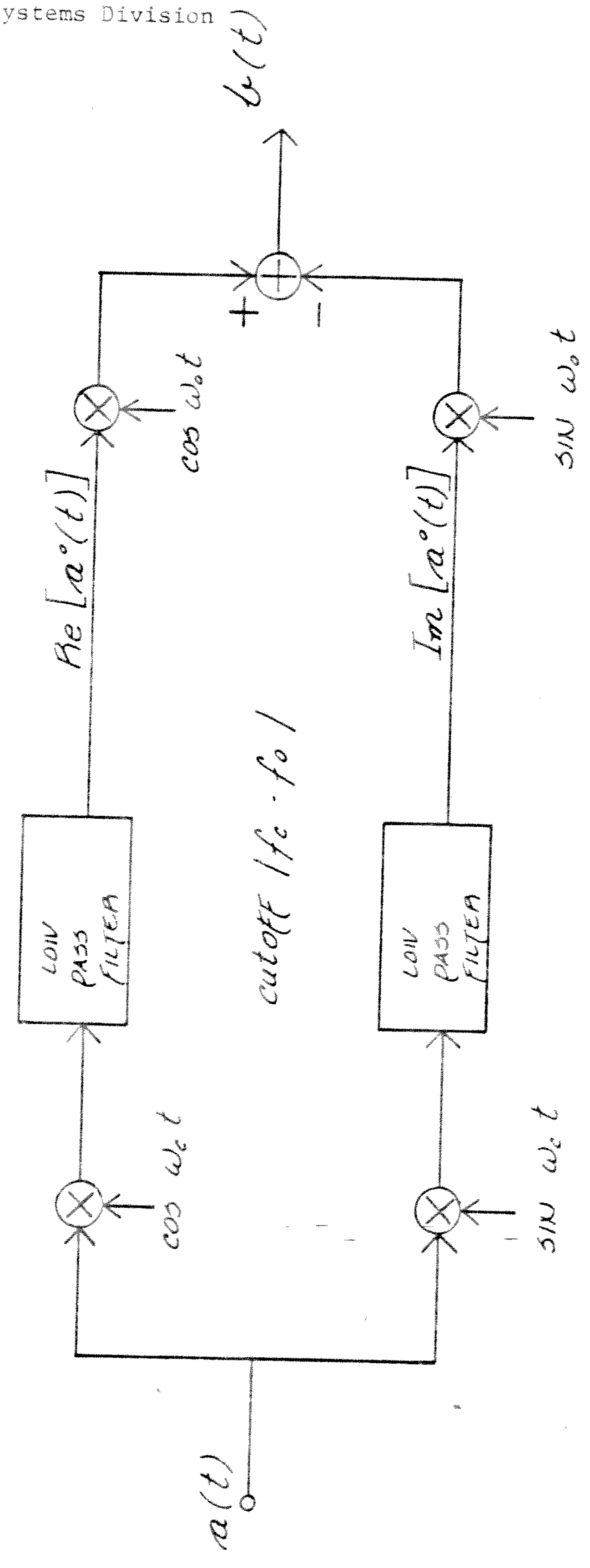


Figure A-5 - Complex Demodulation - Modulacion