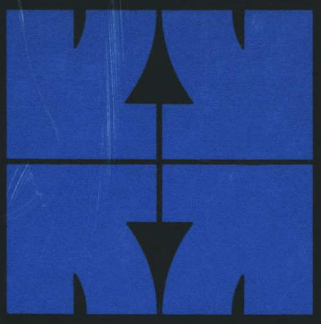
Intelligibility Evaluation of a Simulated 4 kb/s Residual-Excited Linear Prediction Cod

by

Paul Mermelstein

Mamoru Nakatsui



INRS
TÉLÉCOMMUNICATIONS

Institut national de la recherche scientifique

Intelligibility Evaluation of a Simulated 4.8 kb/s Residual-Excited Linear Prediction Coder

by

Paul Mermelstein

Mamoru Nakatsui

INRS-Télécommunications (Université du Québec)

3, Place du Commerce

Ile des Soeurs, Verdun, Québec H3E 1H6

for

Department of Communications, Ottawa

Final Report

Contract No. OSU79-00250

DDS File No. 20SU-36001-9-1396

Nov. 30, 1979 to March 31, 1980

Application of Digital Speech Coding Techniques to Satellite Communication Systems

INRS Report No. 80-05

TABLE OF CONTENTS

ABSTRACT

A residual-excited linear prediction coder transmitting speech at 4.8 kb/s has been evaluated for the intelligibility of the reconstituted signal. Unadjusted intelligibility rates on the Diagnostic Rhyme Test, as measured by inexperienced listeners, were 94.5% for RELP coding only, 91% for RELP coding with 1% transmission bit-error rate, and 84.1% for RELP coding of speech in 10 dB SNR background noise. The results confirm that the RELP technique is significantly more robust to transmission errors and background noise than conventional linear prediction coding.

RESUME

On a évalué l'intelligibilité du signal réconstitué d'un codeur à prédiction linéaire excité par signal résiduel (PLER) qui transmet la parole à 4,8 kilobits/second. Les taux d'intelligibilité non-ajustés selon le "Diagnostic Rhyme Test" pour des sujets naïfs étaient de 94,5% pour codage PLER seul, de 91% pour codage PLER avec un taux d'erreur de transmission en bit de 1%, et de 84,1% pour codage PLER de la parole avec un bruit de fond de 10 dB (rapport signal sur bruit). Les résultats confirment que la méthode PLER est, de façon significative, plus robuste aux erreurs de transmission et au bruit de fond que les codeurs conventionnels à prédiction linéaire.

1. INTRODUCTION

One of the prime technical requirements in applying digital speech compression techniques to airplane to satellite communication is robustness of speech intelligibility to background noise and transmission errors. Speech compression techniques that allow the regeneration of clean speech with adequate quality may not provide acceptable output in the presence of significant background acoustic noise. These narrowband techniques are mostly based on a model of the momentary speech signal that corresponds to one of two exitation conditions, periodic or voiced and aperiodic or turbulent. The techniques can be made more robust by avoiding such binary excitation decisions that can not be made reliably when the speech signal is corrupted by noise. The object of the work reported here was to evaluate to what extent a technique that attempts to overcome these limitation, namely residual-excited linear prediction, in fact preserves intelligibility under these conditions.

Narrowband speech coders allow speech transmission with noticeably degraded quality. Currently 2.4 kb/s is the lowest acceptable transmission rate for codecs of modest complexity that encode individual short frames of the speech signal independently. When the speech signal is not varying rapidly, this frame rate may be reduced without loss of intelligibility, but the requirement for rapid tracking of the spectral shapes in consonantal sounds prevents reduction of the frame rate everywhere. The linear prediction technique is most commonly used today for speech transmission at 2.4 kb/s and generally a slight improvement in quality

results if the bit rate is increased to 4.8 kb/s. Channel vocoders have allowed speech transmission of comparable quality but their implementation today is more costly. Residual-excited linear prediction coders do not allow significant transmission rate reductions below 4.8 kb/s and their speech quality may be somewhat worse than the better LPC implementations. It is the potential robustness to background acoustic noise and transmission errors without requiring special protection bits that recommends a RELP coder for the satellite communication application at hand.

The limited scope of this contract did not allow for the exploration of extensive improvements to the available RELP coding technique. The major novel aspect of the coder used here is the sub-band coding of the residual that allows a reduction in bit-rate from 9.6 kb/s to 4.8 kb/s with but a slight reduction in quality. Our technical objective is to use the evaluation results obtained in the course of this project to guide further design improvements to the RELP coder and thereby further improve speech quality and intelligibility.

The RELP-coder has been simulated on a general-purpose speech processing facility at INRS. It is described in detail in Chapter 2. With the aid of the simulation, speech samples constituting the Diagnostic Rhyme Test were processed under a variety of test conditions. The processed speech data were tape-recorded and presented to listeners for intelligibility evaluation. The details of these evaluation experiments constitute Chapter 3. The overall intelligibility results are discussed in Chapter 4 together with comments on some of the more frequent phonemic

confusions observed. Additional results on the speeh quality attainable with the coder, as measured on phonetically balanced sentences, are given to give the reader a rough idea of the quality attained. This assessment did not form part of the original contract, but is provided merely to supplement the intelligibility results.

2. DESIGN OF THE RELP CODER

2.1 Analysis

The analysis phase of the RELP codec is shown in schematic form in Fig. 2-1 and is essentially that developed by Un and Magill [1]. First of all, the input speech signal s(n) is pre-emphasized by the differencing operation

$$y(n) = s(n) - c_p s(n-1) \qquad (2\text{-}1)$$

where c_p is chosen to be 0.95. This provides about a +6 dB per octave boost to the spectrum of the speech signal thereby reducing its dynamic spectral range. The pre-emphasized speech is then modelled by the linear predictor

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

using the autocorrelation method [2] to extract optimal estimates of the predictor coefficients, a . An eight-pole analysis is used, producing eight LP coefficients a1, a2, ..., a8, per frame. (The analysis frame is 38 milliseconds and the frame is advanced at a rate of 50 times per second. The input speech is digitized at 8000 Hz). The residual signal, r(n), is defined as the difference betwen the original and predicted speech signals, i.e.,

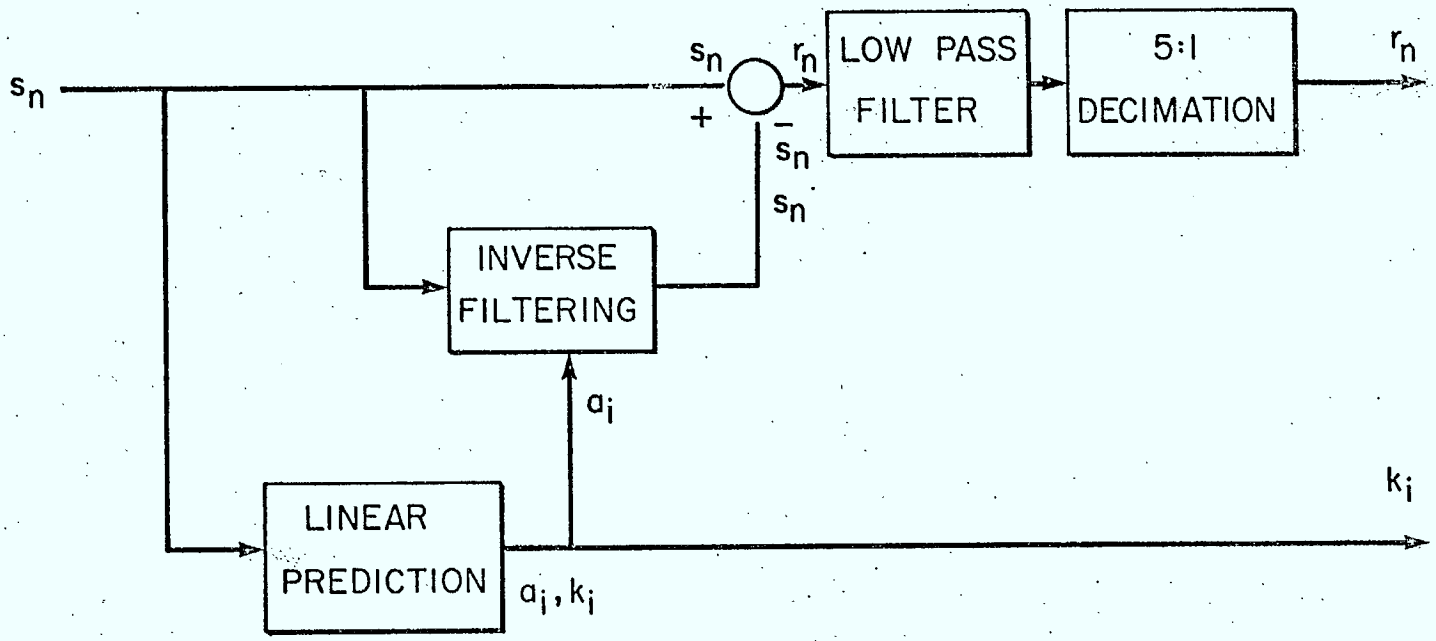


Fig. 2-1 Block diagram of the RELP analyzer. The a_i and k_i represent the LP coefficients and the reflection coefficients, respectively.

$$r(n) = y(n) - \tilde{y}(n) \qquad (2\text{-}3)$$

The residual is conveniently obtained by inverse filtering the pre-emphasized speech signal using

$$r(n) = \sum_{k=0}^{p} b_k y(n-k) \qquad (2\text{-}4)$$

where the coefficients b are related to the a by $b_0 = 1$ and $b_k = -a_k$, $k = 1,\ldots,8$. The residual so extracted resembles the original speech waveform in that it retains the voicing information at the correct place in the waveform. The net saving in transmission results from the fact that (a) the power level of the residual is approximately an order of magnitude less than the original speech signal, and (b) the spectrum of the residual is white and therefore can be approximated at the receiver by a signal which has similar - but not necessarily identical - spectral and temporal properties. For transmission, the residual is low-pass filtered 0 - 800 Hz and decimated 5 to 1. The RMS levels of the 0-800 Hz baseband and 800-4000 Hz high-frequency component are extracted at this point and transmitted along with the predictor coefficients.

The sub-band coding of the residual is carried out as a separate phase. The 800 Hz baseband redidual is first split into two 400 Hz bands using 36-tap quadrature mirror filters. The quadrature mirror filters are designed such that a band splitting arrangement (see Fig. 2-2) will result in perfect reconstruction (aliasing terms cancel) in the absence of quantization or other processing. A passband ripple of the filter is less than 0.2 dB and the minimum stopband attenuation is 33 dB. The 400-800 Hz

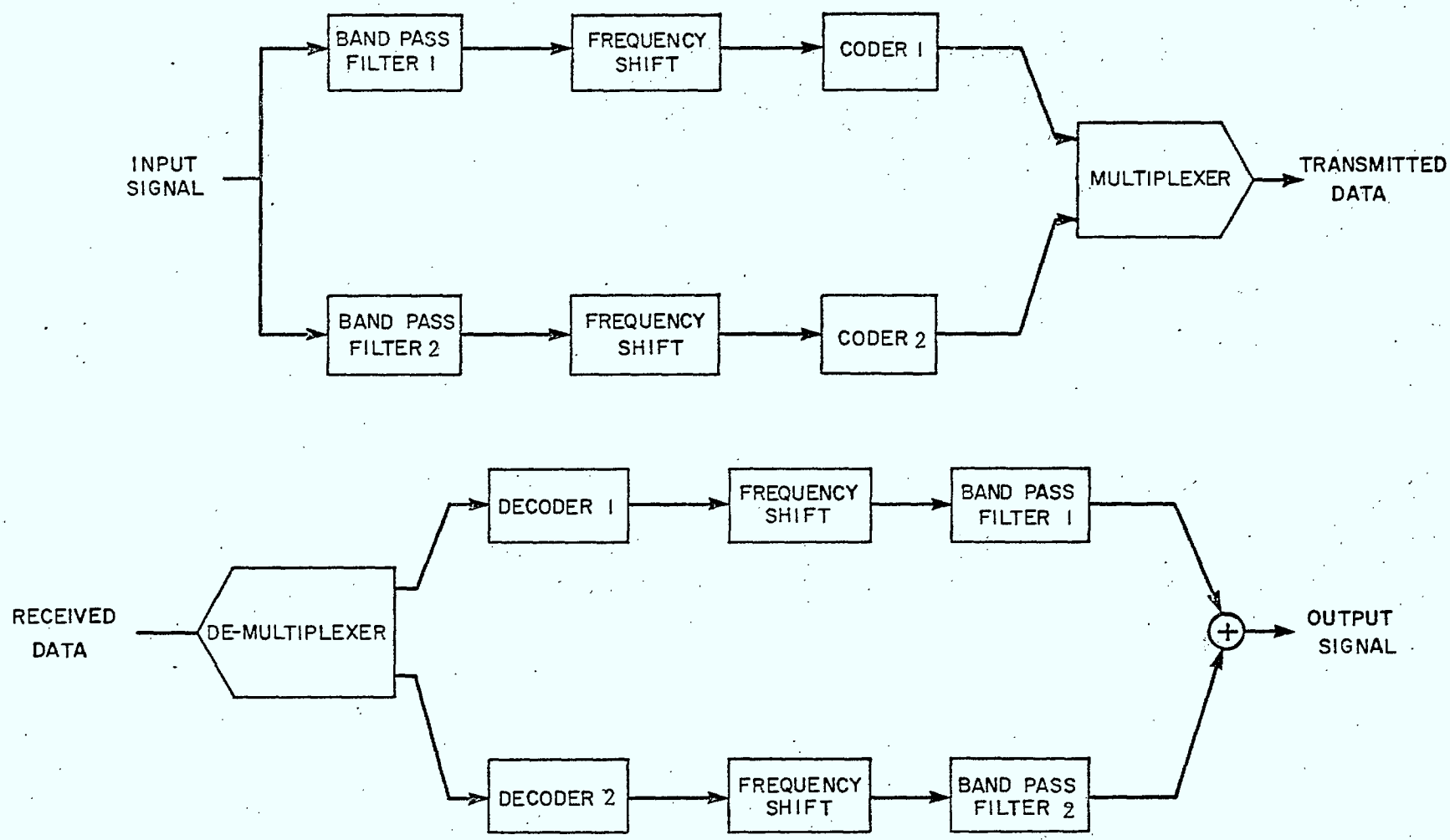


Fig. 2-2 Sub-band coder for the residual.

band is frequency-shifted to 0-400 Hz. Both sub-bands are then down-sampled by a factor of two and then quantized to 2 bits using a one-step adaptive quantizer. The quantization intervals are uniform and scaled by an adaptation parameter that is selected to be 0.85 or 1.9 depending on whether the last sample occupies the two inner or two outer intervals, respectively. Since the effective sampling frequency of these two basebands is now only 800 Hz, this results in a total bit rate of 3200 bits per second. The gain parameters (RMS levels of the baseband and high-frequency component) are allowed 6 and 4 bits, respectively. The eight reflection coefficients are accorded 5,4,4,3,2,2,1 and 1 bits. This represents a total of 32 bits per frame for the spectral and gain information, or 1600 bits per second.

2.2 Synthesis

The 800 Hz baseband residual is reconstructed by upsampling 1:2 and frequency-shifting the sub-band which had originally occupied the range 400-800 Hz. The 0-400 and 400-800 Hz sub-bands are then added to produce the full 800 Hz baseband. This baseband is then used to drive the RELP synthesizer as described below.

The first operation in the synthesizer is to up-sample the residual 1:5, after which the baseband is recovered by interpolation with an 800 Hz low-pass filter. A copy of this baseband is then used to regenerate the high-frequency portion of the residual. This is accomplished with the sequence of operations shown in Fig. 2-3. First, the baseband r(n) is

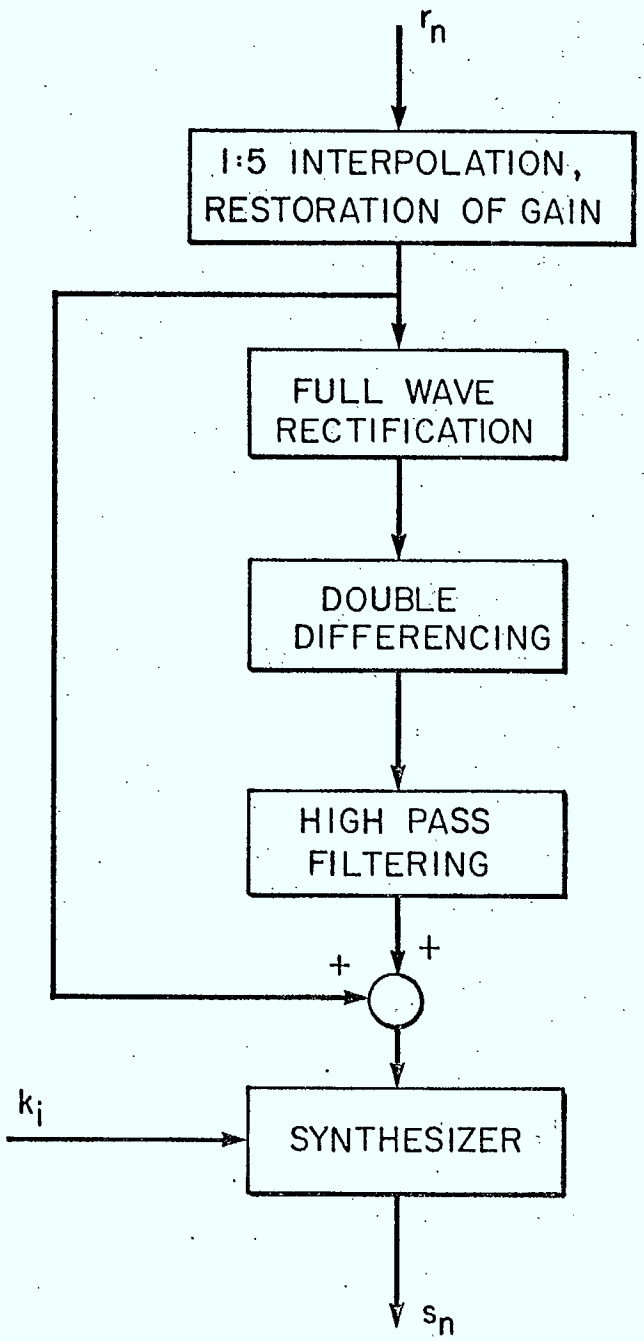


Fig. 2-3 High-frequency regeneration processes used in the RELP synthesizer.

passed through the differencer

$$y(n) = r(n) - 1.4r(n-1) + 0.4\, r(n-2) \qquad (2\text{-}5)$$

It is then passed through a full-wave rectifier to generate the higher harmonics. Since the baseband residual preserves the pitch information of the speech waveform, the harmonics so constructed bear the proper frequency relationships to the fundamental frequency. To restore the residual to its proper spectral balance, the high-frequency component y(n) is flattened by a double-differencing operation [1],

$$r(n) = y(n) - 1.6y(n-1) + 0.64y(n-2) \qquad (2\text{-}6)$$

The HFC is then high-pass filtered from 800 to 4000 Hz and restored to its proper absolute gain level. (The RMS levels of the baseband and high-frequency components are both transmitted on a frame-to-frame basis). The baseband and HFC are then simply added to produce a residual which is both white and possesses a reasonable harmonic structure. The reconstituted signal is then used as the excitation signal for the LP lattice synthesizer. The RELP speech so produced, even in the absence of any quantization of the parameters, shows the hoarse quality typical of RELP speech. The differencing operation applied prior to the full-wave rectification (i.e., Equation 2-5) is found to reduce the hoarseness of the RELP speech at the expense of adding a slight metallic or tonal quality. The coefficients in Equation (2-5) are chosen on the basis of informal listening tests. However, in cases where the residual is heavily quantized (e.g., in a 4.8 kb/s coder), this stage is of lesser importance since

quantization noise in the recovered baseband has an overriding effect on the quality of the high-frequency components of the reconstituted signal.

3. SUBJECTIVE EVALUATION PROCEDURE

The concept of speech quality encompasses the total auditory impression of speech on a listener. As described in Chapters 1 and 2, the RELP technique, which transmits the residual signal instead of pitch and voiced/unvoiced information, is expected to improve not only intelligibility of the transmitted speech but also total quality (including naturalness, speaker recognizability, etc.) over that of existing pitch-excited LPC techniques. In the present study, two methods of evaluating intelligibility and overall quality of the transmitted speech were applied to the RELP system described in Chapter 2. Intelligibility is the most critical performance measure for voice communication systems. The Diagnostic Rhyme Test (DRT) [3] is a currently accepted method for the measurement of speech intelligibility. The DRT test was adopted in the present study to measure intelligibility of the simulated RELP codec under various conditions including frequent channel errors and significant background noise and, in addition, to provide diagnostic information concerning the system performance for various phonetic attributes of syllable-initial consonants.

3.1 Intelligibility Test

One list of 232 words of the Diagnostic Rhyme Test (DRT) provided by Dynastat, Inc., was processed under the following conditions:

1. Band limited high quality digital recording (i.e., uncoded)
2. RELP-coded at 4.8 kbps.
3. RELP-coded at 4.8 kbps with 1% independent binary symmetric bit errors.
4. High quality recording with 10 dB SNR additive white noise.
5. 10 dB SNR additive white noise, RELP coded at 4.8 kbps

Condition (1) served as a control condition against which the intellibility scores of the other conditions were compared.

3.1.1 Speech Material/Digital Data Base

The order of the 232 DRT words (recorded on audio tape at Dynastat, Inc.) had already been randomized in terms of phonetic attributes. This list of 232 words was divided into four sublists, each of which contained 58 words and corresponded to one of four answer sheets. The recorded tape of the DRT words was reproduced with a high-quality tape recorder, low-pass filtered to 3.2 kHz and digitized to 15 bits at a sampling frequency of 8 kHz. No adjustment was made to the loudness levels of the individual words within each sublist because of the apparent uniform vocal effort with which the tape was created. The duration of each spoken word was carefully determined using an interactive audio interface to the computer. At the same time, the average energy of the waveform of each word was measured for later use. These digitized DRT words represented the control condition (1), and also served as input to the RELP simulation program in conditions (2) through (5).

3.1.2. Signal Processing

For conditions (2) and (3), the data base of 232 words was processed by the RELP simulation program described in Chapter 2. No adjustment or optimization of the system parameters of the RELP simulation program was made to the talker of the DRT tape. (The coder parameters had previously been optimized on a speech data base consisting of four sentences spoken by each of five male and four female talkers). One per-cent independent binary symmetric bit errors were introduced by modification of the coder-decoder subprograms of the RELP simulation program. The processed output words were again stored on disk for later testing.

Gaussian white noise was generated with a random number generator and low-passed to 3.2 kHz. This bandlimited white noise was added to the speech data so that a 10 dB average-speech (over 232 words) to average-noise ratio would result. As a result of this SNR definition, the average SNR of individual words ranged from 1.2 dB to 14.6 dB. Although higher intelligibility scores would be expected if the SNR were made uniform across the individual words, the above-mentioned procedure with overall SNR adjustment was chosen because of the continuous vocal effort of the successive words described in 3.1.1. The 232 DRT words with 10 dB SNR additive noise represented condition (4). The same words served as the input data base for the RELP program to produce the stimuli for condition (5).

3.1.3 Test Format

The 20 sublists (5 conditions times 4 sublists) were randomized over the five test conditions listed above, with a pause of approximately 1.5 seconds inserted between successive words. An audio test tape was generated under program control, the audio signal being low-pass filtered to 3.2 kHz on playback. In order to provide a realistic listening situation for conditions (4) and (5) (the conditions incorporating background noise), the pauses between words were filled with Gaussian white noise in condition (4) and with RELP-coded white noise in condition (5). Altogether, 1160 DRT words were presented to ten university students through TDH-39 head-sets in a quiet room. All the listeners were native English speakers and none had previous experience in intelligibility testing. No special training session was provided before the test except for the presentation of a trial sequence of about ten words during which the listening level was adjusted to a comfortable value. The listening level was kept constant throughout the test sessions. Because of the sudden change in speech quality when switching between sublists (each of which represented a different processing condition), two extra DRT words were added to the beginning of each sublist. The judgements on these two words were not included when calculating the intelligibility scores.

3.2 Overall Quality Assessment

A promising single absolute measure of overall speech quality is a subjective speech-to-noise-ratio (SNR) which is derived from pair-comparison tests [4]. To determine this measure the test signal

(i.e. the speech signal processed by the coder to be evaluated) is compared with reference signals corrupted by varying amount of multiplicative white noise. The subjective SNR of the test signal is defined as the SNR of that reference signal which, on the average, is equally preferred to the test signal by groups of listeners. The subjective SNRs of the typical speech waveform coders such as logPCM, ADM with one bit memory and ADPCM with the variable or fixed third order predictor have been evaluated with the method of constant stimuli [4]. The experimental results so obtained provide a useful data base against which newly-developed coders can be compared.

In the present study, an informal test using a modified up-down procedure was carried out to derive the subjective SNR of the RELP-coded speech under condition (2). Four phonetically-balanced sentence spoken by two male and two female speakers were processed by the 4.8 kb/s RELP. The test signal and the reference signal were repeatedly presented to a subject under computer control. The SNR of the reference signal was manually adjusted by the subject by means of typed keyboard input until an equi-preference condition between the test signal and the reference signal was attained. One male experimenter participated as the subject.

4. RESULTS AND DISCUSSION

The Diagnostic Rhyme Test (DRT) yields a gross indication of speech intelligibility of the system as well as providing additional information relating the specific aspects of the system performance to the states of six elementary phonetic attributes of English consontants [3]. In this chapter, the test results are analyzed and summarized with respect to (1) total intelligibility scores and (2) system performance for each state of six binary-valued phonetic attributes. They are compared with reported performance figures for other speech coding systems. In order to further evaluate the system performance, the test results were analyzed with respect to phonemic confusions and compared with published data on perceptual confusions among English consonants [5].

4.1 Intelligibility Scores

As a gross indication of speech intelligibility, total DRT score (percent correct) Pc is calculated by the formula:

$$Pc = 100\ R/T \qquad (4\text{-}1)$$

where R is the number of right answers and T is the total number of items involved. However, Voiers, a proponent of the DRT, has recommended use of an "adjusted percent correct score" (Pa) which compensates for inflated

guessing rates:

$$Pa = 100\ (R - W)/T \qquad (4\text{-}2)$$

where W is the number of "wrong" answers [3]. Both the unadjusted and adjusted scores (Pc and Pa) are given in this paper since both scores have been used in the literature to represent intelligibility [6,7,8]. Furthermore, as will be discussed in 4.3, the adjusted score is not always the better indicator of intelligibility. The following discussion refers to the unadjusted score unless otherwise indicated.

Table 4-1 shows the DRT scores for the five conditions involved in our experimental evaluation of RELP-coded speech. It has been reported that adjusted DRT scores for the original tape are greater than 98% [7]. The lower scores obtained in our control condition (1) may be an artefact of bandlimiting our digital speech data base. Another conributing factor may be that the subjects were not experienced in participating in intelligibility tests, in comparison to the experienced subject pool used by Dynastat, Inc. Since comparison of absolute intelligibility scores obtained at different laboratories is always confounded by differences due to the pool of listeners, prime attention will be paid to the relative intelligibility scores obtained under the five processing conditions used in the present study.

According to Table 4-1, the RELP coding evidently causes approximately 3% decrease in intelligility and a further 3.5% degradation results from the introduction of a 1% transmission error rate. In the presence of

Table 4-1 DRT scores of the RELP; unadjusted(Pc) and adjusted(Pa) scores in per-cent.

Test Conditions	Pc	S.E.	Pa	S.E.
1. High quality data base	97.5	0.27	95.1	0.60
2. Coded	94.5	0.45	89.1	0.83
3. Coded with bit errors(1%)	91.0	0.99	82.7	1.98
4. Noisy data base(10 dB SNR)	89.0	0.86	77.9	1.71
5. Coded noisy data base	84.1	0.66	68.1	1.32

S.E.: Standard errors of the mean

significant background noise, the RELP coding causes a reduction of approximately 5% in the DRT score.

Although no comparable data were found concerning DRT scores for other RELP systems, some reported DRT scores for LPC systems with comparable transmission rates are summarized in Table 4-2. In the table, the data reported by Voiers [8] are DRT scores based on a large set of LPC systems evaluated at his institution over a period of years. The scores quoted are adjusted scores (Pa). The scores reported by Wong and Markel [6] are unadjusted DRT scores (Pc) using a reduced number of DRT words. (Since no numerical scores are given in these reports, all the scores were estimated from the diagrams). Comparing the results of Voiers and of Wong and Markel (Table 4-2) with those of the present study (see Table 4-1), it is clear that the RELP system shows a considerable improvement in total DRT score over typical LPC systems.

No directly comparable data are available in the literature concerning RELP performance for noisy channels and/or background noise. The second line of Table 4-2 shows a normative score of about 87% for a 5% transmission error rate in which all LPC systems allocated a significant number of bits to error protection. Paul has proposed a Speech Envelope Estimation Vocoder which utilizes noise cancelling techniques and operates in the range 2.4 - 8.8 kb/s [9]. The DRT score of his system in the presence of airborne command post noise shows about 10% degradation (87.5% to 78%). However, it is not clear what to make of these figures since the level of background noise is not given and it is not clear whether or not the intelligibity scores have been adjusted for guessing (i.e., Equation

Table 4-2 DRT scores of typical LPC systems.

Reporters(year)	Coder	Transmission Rate(b/s)	Transmission error(%)	DRT score(%) Pc	DRT score(%) Pa	Remarks
Voiers(1979)	LPC	4800	0	93*	86	Normative
	LPC	4800	5	87*	74	score
Wong & Markel(1978)	LPC	5333	0	94	88*	Reduced DRT

* Scores inferred from either the adjusted score Pa or unadjusted score Pc

4-2). The present results (91% with 1% channel errors and 84% in the presence of 10 dB SNR background noise) indicate that the RELP system is acceptable without special protection against noisy environments.

4.2 Diagnostic Scores

The DRT is designed to test for the perceptibility of each of the following elementary phonemic attributes of English consonants (with typical word pair examples):

1. Voicing (VEAL-FEEL)
2. Nasality (NEED-DEED)
3. Sustention (VILL-BILL)
4. Sibilation (JOE-GO)
5. Graveness (BONG-DONG)
6. Compactness (COOP-POOP)

The DRT uses 16 to 18 items, or word pairs, to test the perceptibility of each attribute, and the two states of each attribute (present or absent) are presented equally often. An incorrect response to a DRT item indicates that the listener or system under test has failed to recognize the source state of one of six essentially binary perceptual attributes of English consonant phonemes [3].

Fig. 4-1 shows the diagnostic scores of the RELP system for the above

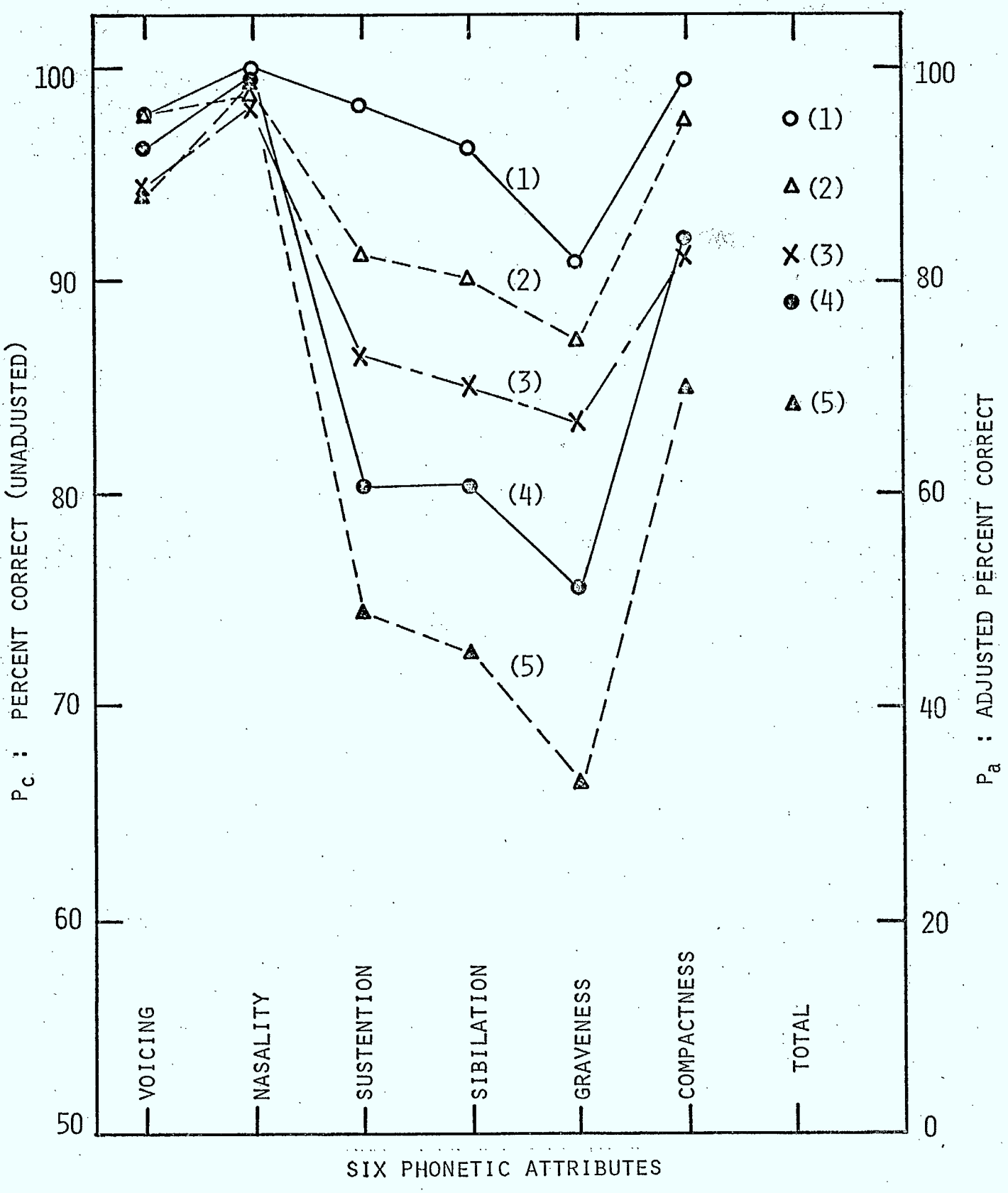


Fig. 4-1 Diagnostic scores for six phonetic attributes.

list of phonetic attributes under the five test conditions. (Hereafter, these scores will be referred to as DRT profiles). It can be seen in Fig. 4-1 that there is considerable consistency in the profiles across test conditions. In other words, even when the test conditions change, no specific degradation is introduced that would significantly affect one feature more than another. This consistency indicates that the RELP coder degrades the intelligibility of the transmitted speech in much the same way as the addition of white noise.

Voiers has reported DRT profiles of the diagnostic scores of typical narrow and medium band digital coders such as linear predictive coders, channel vocoders and CVSDs [8]. Comparing the DRT profile of the RELP in Fig. 4-1 with his DRT profile for a 4.8 kb/s LPC coder, the RELP system shows considerable advantages over conventional LPCs for voicing, nasality and sustention attributes and a disadvantage for the sibilation attribute. All these advantages derive from the utilization of the residual signal as an excitation source. As for the voicing attribute, the distinction between the voiced and unvoiced consonants such as /b/ and /p/ is generally characterized by voice onset timing, i.e., the time interval between the occurence of the stop burst or frication noise and the onset of vocal cord vibration. This information is well preserved in the baseband residual signal, thus avoiding the need to characterize entire frames as voiced or unvoiced as is done in conventional LPC. The transmission of the residual also provides a benefit for consonants which are discriminated by the sustention attribute. For example, the voiced-fricative vs voiced-stop distinction, such as /v/ vs /b/, is characterized by the co-occurence of the sustained voiced excitation (voice bar) and fricative noise for the

former which is not well represented when a system is excited by either buzz or hiss, but not both. The nasality attribute also benefits from residual excitation in that the prominent perceptual cue of nasal consonant, nasal murmur in the range of the base-band characterized by the zero of the spectrum, is well preserved. This is not the case for the pitch-excited (LPC) system since its all-pole constraint cannot successfully model the nasal murmur. Most errors concerning sibilation attribute involve the replacement of the strident fricatives /z/ and /s/ by their mellow counterparts /ð/ and /θ/. Apparently the relative intensity and spectral shape of the frication noise of the strident fricatives is not well preserved by the high-frequency regeneration process used in RELP, as is discussed in detail in 4.3.

4.3 Analysis of Phonemic Confusions

The analysis of phonemic confusions enables one to analyze the system performance in detail by observing whether or not specific perceptual cues for individual consonants are preserved by the system. Table 4-3 shows the eleven most frequent phonemic confusions, for five test conditions. All of these confusions occur with rates greater than twice the average error rate, except for condition (1). In the table we find two general tendencies. First, almost all the confusions are found in all five test conditions. Second, most confusions involve fricatives, affricates and unvoiced velar stops, all of which are characterized to a certain extent by frication or a noise burst.

Table 4-3 List of the eleven most frequent phonemic confusions observed in our DRT test for five conditions and the more frequent confusions observed in Miller and Nilely data with background noise [5]

Spoken	f	θ	s	ʃ	v	ð	z	ʒ	p	t	k	b	d	g	m	n
Condition 1	θ	f	θ	ʃ			ð		t							
Condition 2	θ	f	θ	ʃ			ð		k,t			v,d			n	
Condition 3	θ	f	θ	ʃ			ð		k,t		t	g,v				
Condition 4	θ	f	θ	ʃ	b	z	ð		t	p,k						
Condition 5	θ	f	θ	ʃ			ð		t	p,θ	t	v	g			
Miller,et.al.	θ	f	θ		ð	z,v	ʒ		k	p,k	p,t	g,v	g	d		

Miller and Nicely [5] have reported in detail on perceptual confusions among English consonants and their results have been widely accepted and utilized for many years. In their experiment, listeners were required to select one consonant out of 16 candidates (affricates such as /ʃ/ and /ʒ/ were not included). All of the consonants were followey by the vowel /a/ and the speech signal was band-limited from approximately 200 to 6500 Hz. Though their experimental framework is different from that of the present study, it is interesting to compare their data with the results of the present study. Some of their data for 0, 6 and 12 dB SNR have been rearranged and summarized with respect to the phonemic confusions and total percent correct scores (see Tables 4-3 and 4-4).

The percentage correct scores calculated from the Miller and Nicely data are shown in Table 4-4. Comparing these scores with the DRT scores for 10 dB SNR additive noise shown in Table 4-1, it appears that the unadjusted DRT score is more consistent with the Miller and Nicely results than is the adjusted score. A detailed analysis of the confusion rates shows that there are some phonemic contrasts which give error rates considerably higher than 50%. Since such an error rate would result in an adjusted score of less than 0%, it would appear that the unadjusted score is the more appropriate index of intelligibility. Though a discussion of the "true indicator" of intelligibility is beyond the scope of the present study, the true intelligibility does appear to be closer to the unadjusted score.

The most frequenct confusions observed in Miller and Nicely data are similar to those of this study as shown in Table 4-3. This similarity

Table 4-4 Percent correct scores on consonant identification among 16 English consonants(after Miller and Nicely, 1955).

SNR (dB)	12	6	0
% correct	90.9	83.4	70.7

suggests that the RELP coding degrades the intelligibility of the transmitted speech in much the same way as the additive white noise.

The word pairs observed as exceptions from the above mentioned similarity of the consonant confusions were carefully analyzed with the aid of waveform plots and sound spectrograms. Most of these confusions were found to be related to the labial vs dental opposition in stops and nasals(i.e. /p/ to /t/, /b/ to /d/, and /m/ to /n/). All were followed by the vowel /i/ or /e/, both of which have high second and third formant frequencies. The major acoustic-phonetic cue of these labials is the fast upward transitions of the second and third formants and, in addition, the relatively short voice onset time (for the unvoiced stops). The relative timing of the voicing onset of /p/ and /t/ is well preserved in the RELP-coded speech. There is no confusion between labial/dental pairs which are followed by vowels having low second and third formant frequencies (e.g., /o/ and /æ/). The error rates are only 8.8% for /b/ vs /d/, 7.5% for /m/ vs /n/ and 7% for /p/ vs /t/. The rapid formant transitions of the labials appear to be lost as part of the coding process.

Phonemic confusions concerning the /z/ vs /ð/ and /s/ vs /θ/ pairs contribute significantly to the reduction in the diagnostic score for the sibilation attribute as discussed in 4.2. The strident vs mellow distinction in the fricatives is usually characterized by the relative amplitude and spectral shape of the frication noise and the formant transitions. For the /z/ vs /ð/ pairs, six out of 10 listeners incorrectly identified /z/ under condition (1), i.e., the control (uncoded) condition. Nine out of ten identified /z/ as /ð/ in condition (2). An examination of

the time waveform and spectrograms showed that the /z/ and /ð/ tokens from this particular speaker were very similar. Therefore, the confusibility of these two phonemic contrasts may be a result of idiosyncratic pronunciation habits of that particular talker. Also, the spectral shape of the frication in the uncoded /z/ and /s/, which are usually relatively wide and flat to those of /ð/ and /θ/, are not successfully reproduced in the coded versions.

The somewhat poor performance in preserving the extremely fast formant transitions of stops and the delicate balance in the spectral shapes of fricatives appears to indicate limitations in the performance of the RELP with the current system architecture operating at the significantly reduced data rates. The limitation in intelligibility in these cases is probably due to quantization errors in the base-band coding and spectral inaccuracies resulting from the high frequency regeneration process rather than from representing the spectrum by eight poles at a frame rate of 50 per second.

The main problem with the RELP algorithm is that the reconstructed speech sounds "harsh". At the expense of this harshness, however, the RELP coder preserves very well the overall quality of speech in terms of intelligibility, naturalness and speaker recognizability. In contrast, the speech reconstructed by pitch-excited LPC sounds generally clearer but less natural. Both the merits and disadvantages of the RELP coder have the same origin, i.e., the fact that a portion of the residual is transmitted to the receiver. The robustness of the RELP-coded speech with respect to noise and error degradations is a result of the additional spectral and temporal

information contained in the baseband residual, and the harshness results from the fact that the major part of the spectrum must be regenerated.

4.4 Overall Quality

The overall quality of the RELP-coded speech was assessed with the informal testing described in 3.2. The subjective SNR determined for four sentences is 13 dB SNR. The subjective SNR of the RELP is plotted in Fig. 4-2 together with the subjective SNRs of a variety of waveform coders obtained in a prior study [4]. Overall quality of the 4.8 kb/s RELP is close to that of 32 kb/s log PCM (4 bits quantization at an 8 kHz sampling frequency). This indicates that the RELP technique proposed in this study achieves a bit rate compression factor of 6.7 relative to conventional PCM coding techniques while still preserving similar overall quality. The compression rate is considerablly higher than the typical factors of 3 to 4 achieved by the higher-quality adaptive waveform coders such as APC. However, the associated speech quality is significantly degraded compared to those techniques.

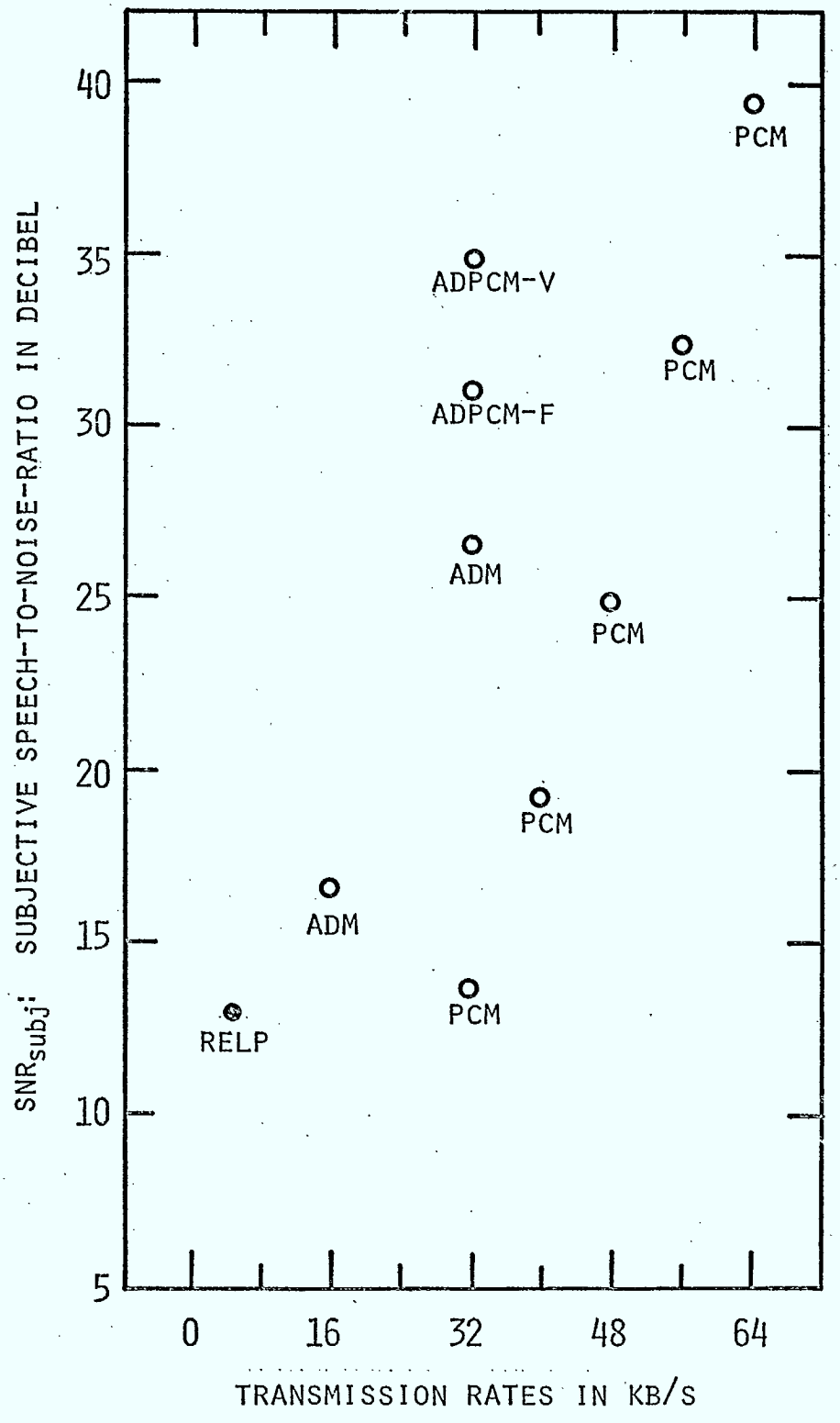


Fig. 4-2 Subjective SNRs of the RELP(●) and of typical waveform coders(○; after Nakatsui, 1980).

5. CONCLUSIONS

The simulated residual-excited linear prediction coder has been evaluated in detail for intelligibility and informally for quality. Intelligibility as measured by the Diagnostic Rhyme Test is 94.5% (unadjusted), 91% for 1% transmission errors and 84.1% in 10 dB average SNR background noise. The performance requirements for the coder were specified as better than 90% for speech with no background noise or transmission errors, better than 85% in 20 dB peak speech to noise ratio, and better than 85% under 1% transmission bit-errors condition. Studies of power distribution in speech [10] allow us to translate the 20 dB peak speech to noise ratio figure to 10 dB mean speech to noise ratio. Although the performance with background noise is slightly below the 85% specification, the difference is not expected to be significant in practice. The other performance requirements are clearly satisfied. The quality of the coder is judged informally to be close to 4 bit log PCM, or near the lower quality limit of communication-quality speech coders. It appears acceptable for field communications but further tests in the actual operating environment are recommended.

Although the RELP design offers advantages of simplicity over LPC coders in that the pitch-tracking component is eliminated, no quanititative measure of circuit complexity can be made at this time. Examination of the hardware requirements for the RELP coder will form the next stage of the study. On the basis of the intelligibility evaluation, two areas of RELP coder design have be noted for further study as likely to lead to

intelligibility improvements. The phase relationship between the regenerated high frequency components and the transmitted low-frequency component of the residual signal is not well known. We suspect that the phase incoherence in the reconstituted signal contributes to the difficulty in perceiving rapid formant transitions as well as to the harsh quality of the reconstituted signal. It may be useful to introduce additional random noise into the reconstructed residual in addition to the noise generated as a result of quantization of the baseband. Such additional noise may improve the intelligibility of the strident fricatives.

ACKNOWLEDGEMENTS

The simulation of the RELP coder is the work of Dale Stevenson of Bell-Northern Research. The design of the sub-band filters and adaptive quantizer is due to Peter Kabal. The subjective listening experiments were conducted by Elliott Dainow. The assistance of these colleagues in the execution of this study is much appreciated.

REFERENCES

[1] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," IEEE Trans. Vol. COM-23, pp. 1466-1474, Dec. 1975.

[2] J. Markel and A. Gray, "Linear Prediction of Speech," Springer-Verlag, Berlin 1976.

[3] W. D. Voiers, "Intelligibility testing at Dynastat: the Diagnostic Rhyme Test," private document by Dynastat, Inc., Austin, Texas, 1978.

[4] M. Nakatsui and P. Mermelstein, "A unified subjective measure of speech quality for digital coders," The 10th Intern. Conf. on Acoust., Sydney, Australia, July 1980 (in print).

[5] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Amer., Vol. 27, pp. 301-315, 1955.

[6] D. Y. Wong and J. D. Markel, "An intelligibility evaluation of several linear prediction vocoder modifications," Contract Report (No. MDA904-76-C-0259), Signal Technology, Inc., Santa Barbara, California, March 1978.

[7] S. Maitra and C. R. Davis, "A speech digitaizer at 2400 bits/s,"

IEEE Trans. Vol. ASSP-27, pp. 729-733, Dec. 1979.

[8] W. D. Voiers, "Diagnostic evaluation of speech coding systems," Final Report (Contract No. DCA100-77-C-003), Dynastat, Inc., Austin, Texas, Sept. 1978.

[9] D. B. Paul, "A robust vocoder with pitch-adaptive spectral envelope estimation and an integrated maximum-likelihood pitch estimation," Record of the 1979 IEEE Intern. Conf. on Acoust. Speech and Signal Processing, pp. 64-68, Washington, D. C., April 1979.

[10] D. L. Richards, "Telecommunication by Speech," John Wiley & Sons, N. Y. 1973, p. 441.

MERMELSTEIN, PAUL.
--Intelligibility evaluation of a simulated 4.8 kb/s residual-excited linear prediction coder.

Université du Québec
Institut national de la recherche scientifique
INRS-Télécommunications

a/s Bell Northern limitee
3, place du Commerce
Verdun, Québec
H3E 1H7