IEEE Canadian Communications and Power Conference

October 16, 1980
Montreal, Quebec

A COMPARATIVE STUDY OF DIGITAL CODING TECHNIQUES AT 16 KB/S AND BELOW

J. Turner
Bell-Northern Research
Verdun, Quebec H3E 1H6

D.C. Stevenson
Bell-Northern Research
Verdun, Quebec H3E 1H6

P. Kabal
INRS-Telecomm., University of Quebec
Verdun, Quebec H3E 1H6
and
Electrical Eng., McGill University
Montreal, Quebec H3A 2A7

P. Mermelstein
Bell-Northern Research
Verdun, Quebec H3E 1H6
and
INRS Telecomm., University of Quebec
Verdun, Quebec H3E 1H6

Abstract

Four recently-developed coding techniques are surveyed: Adaptive Predictive Coding (APC), Adaptive Transform Coding (ATC), Residual-Excited Linear Predictive Coding (RELP) and Sub-band Coding (SBC). These techniques have the common property of trying to minimize the perceived distortion by exploiting characteristics of both the speech signal and the listener. APC and ATC use adaptive techniques to track the short-time statistics of the input speech. At 16 kb/s, the reconstructed speech quality from these coders approaches that of telephone toll standards (56 kb/s companded PCM). Sub-band coding allows the quantization noise in each of several frequency bands to follow the short-time speech energy. Residual-excited Linear Predictive Coding transmits only a low-frequency base-band and uses that component to regenerate the higher speech frequencies. Both RELP and SBC produce speech whose quality is acceptable for many communications applications at 9.6 kb/s. The results of subjective preference testing of these coders are presented and their relative complexities are discussed.

Introduction

The slowly varying short-term energy and spectral envelope of speech signals permit bit rate compression through coding. The actual rate of energy and spectral change depends on the type of speech sounds. Adaptive quantization and adaptive prediction take advantage of these properties by reducing, in a reversible way, the variation of the speech signal being coded. Instead of direct quantization of the time samples of the speech signal as in waveform coders such as PCM, these adaptive procedures generate a normalized residual signal which is quantized and reconstructed into a speech signal. For the same perceptual quality, this new signal can be quantized at a lower bit rate. Time domain waveform coding techniques employ these two procedures in various fashions to obtain the desired trade-off between bit rate, coded speech quality and coder cost.

Time Domain Waveform Coders

Adaptive quantization essentially matches the size of the quantizer levels to the short-term signal energy. Normalizing the signal before quantization and rescaling after decoding allows a quantizer with fixed levels and a limited dynamic range to be used. The energy normalization procedure can either operate on blocks of speech samples or it can operate on a sample-by-sample basis by calculating a running estimate of the signal energy.

In adaptive prediction methods, a prediction operator combines weighted previous signal samples to estimate the current signal amplitude. This method removes the short-term correlation from the input signal. When the prediction coefficients are optimally chosen, the spectral envelope of the resulting residual signal is approximately flat. Furthermore, the variance of the residual signal is less than that of the original speech signal. The predictor coefficients are determined on the basis of speech statistics from large populations of talkers. Simple waveform coders such as Adaptive Differential PCM (ADPCM) typically use fewer than four fixed predictor terms. For high quality speech, the transmission rate must be above 24 kb/s.

The statistics of the input signal vary from speaker to speaker and also within an utterance from a single speaker. In order to better remove the short-term waveform correlations for non-stationary signals, adaptive prediction methods are used. When the prediction coefficients are updated at intervals of around 30 milliseconds, up to 10 dB SNR improvement can be obtained for most types of speech sounds and talker characteristics. For speech band-limited to 4 kHz, the number of prediction terms typically used is around eight, with more terms producing only a small additional improvement. Neither adaptive quantization nor adaptive prediction are speech specific; they are also applicable to other types of input signals such as data modem signals.

The transmission rate for speech can be further reduced by taking into account the characteristics of the speech generation and perception mechanisms. For instance, English speech can be adequately characterized as either voiced or unvoiced for many coding applications. Voiced speech sounds are characterized by the vocal tract resonances excited at a rate equal to the fundamental (pitch) frequency. Unvoiced speech sounds can be reasonably well described by random noise with appropriate spectral shaping. The pitch period of voiced speech is typically at least an order of magnitude longer than the time spanned by the predictor. In order to remove the pitch redundancy, a different predictor structure which incorporates pitch period tracking must be implemented.

The coefficients from the adaptive predictor together with the pitch-period descriptors constitute a very compact description of the speech signal. For

instance, Linear Predictive Coding (LPC) techniques describe the residual waveform only in terms of the voicing characteristics and the pitch period. At bit rates of 2.4 kb/s, LPC produces highly intelligible speech, albeit with a somewhat unnatural quality. Residual-Excited Linear Prediction (RELP) improves the quality over that of LPC by sending part of the residual signal. At a rate of 9.6 kb/s, RELP produces speech which is more natural than LPC and is also more robust to speaker variations and background noise. It employs an adaptive predictor to determine the residual signal and an adaptive quantizer to encode it. Since RELP does not use a pitch extractor, pitch-tracking errors are avoided. Rather, the low frequency component of the residual is assumed to adequately contain the necessary voicing information. Since the harmonics of the pitch frequency extend through the entire signal bandwidth, the high frequency regeneration procedure can be used to restore the missing portion of the spectrum. In practice, the reconstructed high-frequency component, while having the correct harmonic structure, lacks the short-time phase structure necessary for high quality speech.

Adaptive Predictive Coding (APC) also uses both adaptive prediction and adaptive quantization, but unlike RELP, transmits the full residual signal. A schematic diagram of an APC coder is shown in Fig. 1.

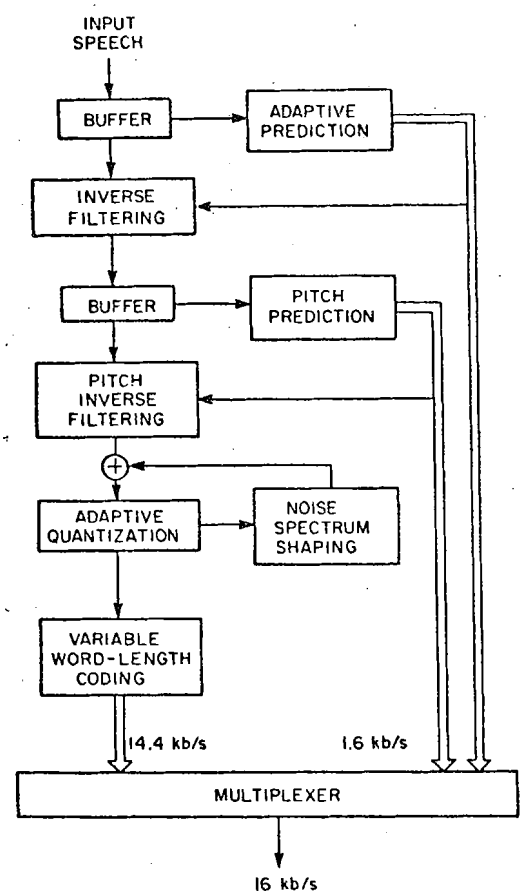

Fig. 1. Diagram of an APC Coder.

In APC, the objective is to produce minimum perceptual distortion rather than maximum objective SNR. This is achieved by adjusting the shape of the coding noise spectrum so that the speech spectrum tends to mask the distortion. APC also uses a variable length code word to represent the quantization intervals. Short code words are used for the most frequently occurring (low amplitude) levels while longer code words are used for the others. This results in a reduced average bit rate but requires that the bit stream be buffered before transmission over a fixed rate channel.

Frequency Domain Waveform Coders

The short-time statistics of the input signal can also be exploited by frequency-domain techniques. In its simplest form, a frequency domain coder such as a Sub-band coder divides the input spectrum into a small number of frequency bands and codes each band separately. The digitizer for each sub-band adapts to the short-term energy in that band. Bits can be assigned to each band in accordance with the signal-to-noise ratio desired in that band. Since the lower frequency bands are found to be perceptually more important than the high frequency bands, fewer bits are used to code the high frequency bands. Additional benefits accrue since spectral shaping of the quantization noise follows the time-varying spectral distribution of the speech and is therefore masked more effectively by the speech signal.

Adaptive Transform Coding (Fig. 2) carries this process further, effectively increasing the number of frequency bands to 128 or 256. The quantization

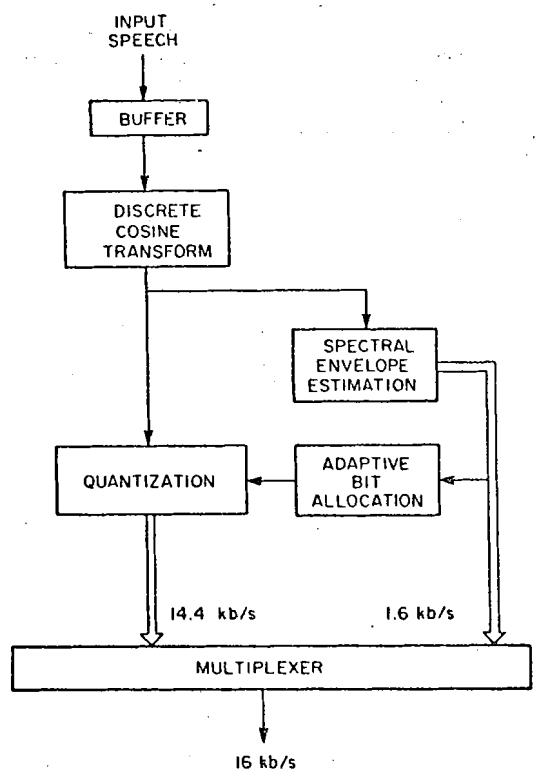

Fig. 2. Diagram of an ATC Coder.

adaptation strategy is based on the short term spectral envelope which is obtained from the optimal predictor coefficients as in conventional time domain techniques. The quantizer for each frequency point is scaled by the amplitude of the power spectrum at that frequency. Further improvement can be obtained by modifying the spectral envelope to allow the high amplitude harmonics to be reconstucted more faithfully than the low amplitude harmonics.

In order for the coding process to be invertible, the spectral envelope shape is sent as side information to the receiver. The coding of the frequency domain samples uses dynamic bit allocation: the total number of bits available for a block of data is allocated among the frequency samples in such a way as to minimize the overall distortion. Noise shaping can also be implemented in ATC by having the allowable distortion vary with frequency. ATC thus is in many ways the frequency domain dual to APC: both schemes take advantage of the same properties of the speech waveform and the human perceptual apparatus in order to make effective coding gains.

Quality Considerations

The four coders described above - APC, RELP, ATC and Sub-band - use a variety of methods to extract and code predictable properties of the speech signal. Because they do so in different ways, each coder creates different types of distortions in the reconstructed speech signal and demonstrates a different senstivity to decreasing bit transmission rate. Subjective tests were carried out to rate the quality of the coded speech at various simulated transmission rates for these coding schemes.

Our version of APC combines features of the coders proposed by Atal and Schroeder [1] and Makhoul and Berouti [2]. At 9.6 kb/s, centre clipping and run-length coding of the residual are used. The RELP implementation is similar to Un and Magill [3] but with an additional double difference stage before rectification of the baseband. This modification trades off hoarseness for a sort of metallic sheen and improves the subjective quality by approximately 0.5 on the log PCM scale. At the lower rate of 4.8 kb/s, the residual can be coded using a sub-band technique to reduce the quantization noise in the received residual and thereby improve quality [4][5].

The Sub-band coder uses the frequency bands proposed by Crochiere, with gaps in the spectrum when coding at 9.6 kb/s [6][7]. At 16 kb/s, a quadrature mirror filter arrangement produces similar results [8][9]. The ATC coder [10] is similar to that of Tribolet and Crochiere [11] in that it uses linear prediction to determine the spectral envelope. Additional block boundary smoothing was also found to be beneficial.

Several phonetically-balanced sentences from the Harvard Lists [12] were coded at 9.6 kb/s and 16 kb/s for both male and female speakers. (The different coders were tested at different times; however, the test sentences were drawn from the same speech data base and the testing procedure was identical for all four coders). The quality was rated by groups of naive listeners over TDH-39 headphones in a quiet listening environment. The subjects rated the quality of each coder against 3, 4, 5, 6 and 7 bit log PCM codings of the same sentences. The equi-preference position on the log PCM scale was determined by the 50% preference value of the least mean square fit to the subjects° preference judgments [13]. These points on the log PCM scale are shown below in Table I.

Additional testing of the coders was carried out in which the coded speech was compared against speech samples which had been degraded by various amounts of multiplicative noise [14]. The subjective SNRs determined in this fashion predict a preference ranking for coder quality that is in general agreement with the results cited in Table I.

TABLE I

EQUI-PREFERENCE POINTS ON LOG-PCM SCALE

	Bit Rate	
	9.6 kb/s	16 kb/s
APC	(4.0)*	6.5
RELP	4.5	(4.8)
ATC	5.4	6.2
SBC	3.9	(5.8)

* Figures enclosed in parentheses were estimated from informal listening tests

Although the types of distortions introduced by these four coders are not the same, the equi-preference results shown in Table I allow one to rank these coders in terms of overall quality. Using these results and other published estimates of coder quality (e.g., Flanagan et al., [15]), quality curves over the entire bit transmission rate can be estimated as shown in Fig. 3. Also illustrated is the

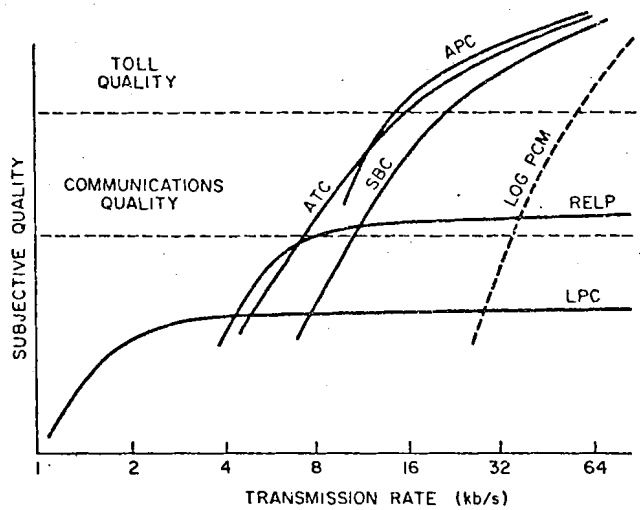

Fig. 3 Subjective quality as a function of transmission rate.

log PCM quality curve which approaches an asymptotic increase in quality of 6 dB/bit.

The limiting quality of the RELP speech is determined by the procedure used to regenerate the high frequency harmonics. Since the RELP coder shown in Fig. 3 assumes a fixed baseband (0-800 Hz), no significant quality gains can be achieved above 9.6 kb/s. The limiting quality of the RELP speech is therefore determined by the procedure used to regenerate the high frequency harmonics. Neither APC nor ATC demonstrate such an asymptote in quality. When none of the transmitted parameters is quantized, these coders reproduce the input speech signal very closely. For this reason the quality of the output speech increases as a monotonic function of the bit rate. If the frequency bands used in the Sub-band coder (SBC) completely span the frequency spectrum, the quality curve for the SBC coder is limited only by the filtering noise. However, if one or more frequency bands are omitted, then the ultimate quality of the Sub-band coder is limited by the coding scheme itself.

Coder Complexity

Many issues must be considered when choosing a coder for a given application. Coder complexity, and therefore cost, is currently an important factor in the design of a speech coder. In applications where transmission costs are high or the transmission facility is a scarce resource, a complex coder may be warranted if it more efficiently uses the medium. In general, the transmission bit rate can only be reduced by substantially increasing the coder complexity.

The quality curves shown in Fig. 3 show that APC and ATC produce similar quality speech. At 9.6 kb/s, both of these coders produce speech which is superior to both RELP and Sub-band. However, they do so at a considerable increase in coder complexity. Fig. 4 shows the estimated relative complexity of the four coders

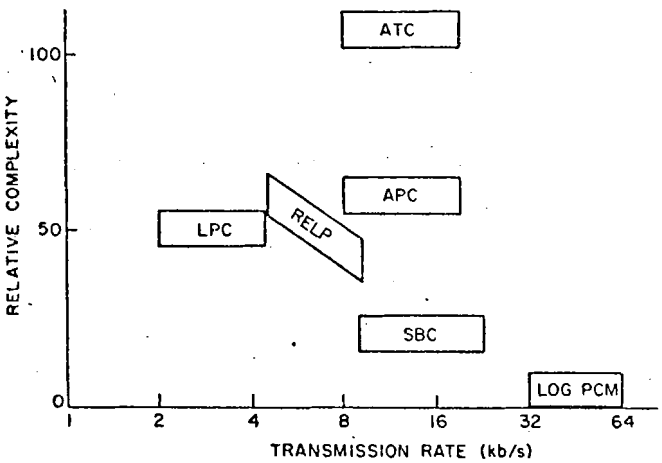

Fig. 4 Coder complexity for various transmission rates.

as a function of bit rate. LPC and log PCM are also shown for purposes of comparison. The complexity of RELP increases slightly at low bit rates (symbolized by the slanted bar in Fig. 4), reflecting the incorporation of sub-band coding of the baseband.

Commercially available IC implementations include simple time domain coders such as Adaptive Delta Modulation. Sub-band coding is amenable to hybrid sampled data and/or digital IC technologies. Low bit rate coders such as LPC are available as microprocessor based units augmented with dedicated hardware processing functions at single quantity costs in excess of $10,000.

Recent advances in IC technology suggest that fairly complicated algorithms such as APC conceivably could be implemented on a single VLSI chip within the next five years. Even the more complicated schemes such as ATC may become practicable with advances in IC technology. For instance, with a single chip FFT processor, an ATC coder could be implemented with only a small number of VLSI chips. Thus we can expect rapidly decreasing costs for the more complex speech coders, making them practical alternatives to the broadband coders in use today.

References

1. B.S. Atal and M.R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criterion", IEEE Trans. ASSP, Vol. ASSP-27, No. 3, June 1979, 247-254.

2. J. Makhoul and M. Berouti, "Adaptive Noise Spectral Shaping and Entropy Coding in Predictive Coding of Speech", IEEE Trans. ASSP, Vol. ASSP-27, No. 1, Feb. 1979, 63-73.

3. C-K Un and D.T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 kbits/s", IEEE Trans. Commun., Vol. COM-23, No. 12, Dec. 1975, 1466-1474.

4. D.J. Esteban, C. Galand, D. Mauduit, and J. Menez, "A 4800 BPS Voice Excited Predictive Coder (VEPC) based on Improved Baseband/Sub-band Filters", IEEE ICASSP, Wash., D.C., April 1979, 975-979.

5. P. Mermelstein and M. Nakatsui, "Intelligibility Evaluation of a Simulated 4.8 kb/s Residual-Excited Linear Prediction Coder", INRS-Telecomm. Technical Report 80-05, April 1980.

6. R.E. Crochiere, S.A. Webber and J.L. Flanagan, "Digital Coding of Speech in Sub-bands", BSTJ, Vol. 55, No. 8, Oct. 1976, 1069-1085.

7. R.E. Crochiere, "On the Design of Sub-band Coders for Low-Bit-Rate Speech Communication", BSTJ, Vol. 56, No. 5, May-June 1977, 747-791.

8. D.J. Esteban and C. Galand, "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes", IEEE ICASSP, Hartford, Conn., May 1977, 191-195.

9. A. Roset, "Application des filtres miroirs à un procédé de codage par découpage en sous-bandes", INRS-Telecomm. Technical Report 80-03, Jan. 1980.

10. D.G. Sloan, "Adaptive Transform Coding of Speech," INRS-Telecomm. Technical Report 79-05, June 1979.

11. J.M. Tribolet and R.E. Crochiere, "Frequency Domain Coding of Speech", IEEE Trans. ASSP, Vol. ASSP-27, No. 5, Oct. 1979, 512-530.

12. "IEEE Recommended Practice for Speech Quality Measurements", IEEE Trans. ASSP, Vol. ASSP-17, Sept. 1969, 227-276.

13. J.D. Finney, Probit Analysis, Cambridge: Cambridge University Press, 1964.

14. M. Nakatsui, "Subjective speech-to-noise ratio as a single absolute performance measure for digital waveform coders", J. Acoust. Soc. Am. Suppl. 1, Vol. 67, Spring 1980 (abstr.).

15. J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E. Crochiere, N.S. Jayant, and J.M. Tribolet, "Speech Coding", IEEE Trans. Commun., Vol. COM-27, No. 4, April 1979, 710-737.