EFFICIENT CODING OF THE PREDICTION RESIDUAL

LEGAND L. BURGE, JR.

DEPARTMENT OF ELECTRICAL ENGINEERING
UNITED STATES AIR FORCE ACADEMY, CO

ABSTRACT

This paper presents an efficient method of coding the prediction residual using the technique of sub-band coding designed at the bit rate of 9600 bits/second. The energy of the prediction residual is used to distribute the bit allocation by sub-bands such that the perceptual criteria is enhanced by transitional information within the phoneme connections of speech by a technique that weights the energy based on a normalization factor. A three tier phoneme classification is derived from an energy study of the phonemes for the prediction residual. With this it is shown that speech intelligibility is enhanced in the coding scheme. An adequate indication of coder quality is described using various types of signal-to-noise ratios.

INTRODUCTION

Recently considerable interest has been given to methods of digital analysis and synthesis of speech assuming a basic model. This model of the speech waveform is assumed to be a linear time invariant system which responds to a periodic or noiselike excitation. This linear time invariant system represents the vocal tract. If the vocal tract is assumed to be fixed, then the output of the system is a convolution between the excitation and vocal tract transfer function.

A method that has proven to be efficient for encoding the speechwave is linear prediction (9). The linear predictive encoder was developed to improve the channel vocoder voice quality and intelligibility. The linear predictive filter is expressed as the equation to describe the frequency response of the vocal tract system. Its function is to decompose the speech into two waveforms. One waveform represents the parameters that are time varying such as predictor coefficients, partial correlation coefficients and other parameters that represent the formant frequency characteristics. The other waveform is the prediction residual. The residual is the difference between the actual and predicted speech signals. Several authors have investigated the coding aspects of the prediction residual which the author has documented in Reference (10).

In this paper the basis of coding the prediction residual is discussed. The classification of phonemes into energy aggregations is presented. The bit allocation scheme is discussed using the energy groupings. Performance measurements are presented as an indication of coder quality.

BASIS OF CODING THE PREDICTION RESIDUAL

A coding method is presented to perceptually enhance the speech. The method uses the basis of sub-band coding (SBC) for coding the prediction residual (3)(7). Besides SBC being conceptually simple, it has the additional advantage that each sub-band is quantized spearately and each band contains its own distortion. It should be pointed out that the input to the sub-band coder is the residual signal rather than the speech signal.

The spectrum of the signal is used for calculation of the energy. The energy can be represented by

$$E = \frac{1}{N} \sum_{k=0}^{N-1} |E_f(k)|^2 \qquad (A)$$

where $E_f(k)$ corresponds to the discrete Fourier transform (DFT) coefficients of the prediction residual signal $e_f(k)$, which can be computed by using the fast Fourier transform (FFT) algorithm.

Equation (A) is applied to the prediction residual to compute the energy. The spectrum of the prediction residual is partioned into four sub-bands as stated before. Using (A), the energies in each sub-band can be expressed by

$$E_n = \frac{1}{N} \sum_{k=0}^{N-1} |E_{fn}(k)|^2 \quad n=1,2,3,4 \qquad (B)$$

where $E_{fn}(k)$ is the DFT coefficient of the signal corresponding to the nth sub-band.

Now the total energy can be expressed by

$$E_T = \sum_{n=1}^{4} E_n \qquad (C)$$

Among speech sounds, E_T has wide variance. Previous researchers have not studied the variations in E_T of the speech sounds for each prediction residual. This aspect is discussed in the next section.

Energy Distribution

This section gives the results on the energy data for phonemes.* The goal of the energy study is to distinguish between vowels, nasals and noisy sounds. This data is used to determine the bit distribution in the coding algorithm. From this analysis, it is shown that the energy of the prediction residual divides the phonemes into classes by phonemic aggregations. The phonemes can be grouped into three classes, namely high energy, low energy, and noise groups. The high energy group includes the vowels and dipthongs. The plosive, fricative and unvoiced phonemes make up the noise group. The low energy group is composed of glides and nasals. It follows that an ideal excitation signal for speech would enhance perception by including a three-source model rather than the conventional two-source model (10). It is well known that with simple LPC methods, the excitation function is a set of periodic pulses or random noises which can be identified as high or low energy excitation functions (9). The three-source model would include a source for vowels, a source for nasals and glides, and a source for fricatives. This is the result of the phoneme energy study of the prediction residual. Further results have been documented in Reference (10).

For the sub-band coding, threshold values need to be computed for each band. Also, each energy group has to be divided into four subgroups corresponding to the four sub-bands. Let E_{in} be the normalized signal energy in the nth frequency band corresponding to the phoneme that is in the ith energy group. This is explicitly shown in Table I. For example, E_{12} represents the energy in the second frequency band corresponding to the high energy phoneme (first energy group).

The threshold values for E_{in} (referred as E^T_{in}) in Table I will now be established. For clarity, let the energy in the nth frequency band be represented by (see (B))

$$B_{in} = E_n \tag{D}$$

TABLE I.
SYMBOLIC REPRESENTATION OF ENERGY DISTRIBUTION

Energy	Frequency Band 1	2	3	4
H	E_{11}	E_{12}	E_{13}	E_{14}
L	E_{21}	E_{22}	E_{23}	E_{24}
N	E_{31}	E_{32}	E_{33}	E_{34}

corresponding to the phoneme that is in the ith energy group. To make the classification speaker independent, the B_{in} in (D) has to be normalized by E_T given in (C). Let

$$E^T_{in} = \frac{B_{in}}{E_T} \quad , \quad \begin{matrix} i = 1, 2, 3; \\ n = 1, 2, 3, 4 \end{matrix} \tag{E}$$

From this, it is clear that

$$E^T_{in} \leq 1.0 \quad \begin{matrix} i = 1, 2, 3; \\ n = 1, 2, 3, 4 \end{matrix} \tag{F}$$

As before, E^T_{in} in (D) are tabulated for i=1, 2, 3 and n=1, 2, 3, 4. The breaks are established from this tabulation and the threshold values are obtained from these breaks. These are tabulated in Reference (10). The array in Table I will be referred hereafter as energy threshold matrix. This matrix will be used in computing the bit allocation scheme, which is discussed in the next section.

*All data referred to but not included in this paper are available in the author's PhD thesis which is available through the Oklahoma State University library (see reference 10).

The Algorithm

In this section, the bit allocation scheme is discussed using the energy groupings established. In symbolic form, the bit distribution is shown in Table II for a three energy level -- four sub-band coder, where the rows correspond to the energy levels and the columns correspond to a particular frequency band. For example, k_{23} bits per sample assigned for the second energy (LE) band and the third frequency band.

TABLE II
SYMBOLIC REPRESENTATION OF BIT DISTRIBUTION

	Frequency Band 1	2	3	4
High Energy (H)	k_{11}	k_{12}	k_{13}	k_{14}
Low Energy (L)	k_{21}	k_{22}	k_{23}	k_{24}
Noise (N)	k_{31}	k_{32}	k_{33}	k_{34}

The bits are allocated by the empirical formula

$$k_{ij} = \log_2\left(1 + \frac{E_{ij}}{\sigma_i}\right) \quad , \quad \begin{matrix} i = 1, 2, 3; \\ j = 1, 2, 3, 4 \end{matrix} \tag{G}$$

where E_{ij} is the energy from Table II and σ_i is a normalization factor determined from the constraint

$$\sum_{j=1}^{4} k_{ij} N_j = C \quad , \quad i = 1, 2, 3 \tag{H}$$

with N_j, j = 1, 2, 3, 4, being the number of samples in each band after decimation. The value of C is equal to the total number of bits/frame minus the number of sync bits. Combining (G) and (H) it follows that

$$\sum_{j=1}^{4} N_j \left(\log_2\left(1 + \frac{E_{ij}}{\sigma_i}\right)\right) = C \quad , \quad i = 1,2,3 \tag{I}$$

where the normalization factor, σ_i, can be determined from (I). Equations (G), (H), and (I) define the algorithm.

The normalization factor is included to take into consideration the perceptual aspects of the signal. It is used as a weighting factor for transitional cueing (6). To compute the normalization factor properly for coding the residual signal, a bit matrix is chosen. The bit distribution is selected that is based on perceptual concepts. This matrix will be referred to as an a priori bit matrix. In addition to perceptual concepts, the a priori bit matrix is selected such that the bit rate is 9600 bits/second for the sub-bands. The matrix shown in Table II, where the entries will now be referred to as k^A_{ij} to denote the a priori values.

When the energy of the speech sound is determined to be high enough, the energy threshold selects the energy matrix (from Table I) and a priori bit values (from Table II). These are used to calculate the normalization factor from (G), and

$$\sigma_j = \frac{E^T_{ij}}{(2k^A_{ij})-1} \qquad \begin{matrix} i = 1, 2, 3 \\ j = 1, 2, 3, 4 \end{matrix} \qquad (J)$$

where E^T_{ij} is the energy obtained from threshold matrix and k^A_{ij} is obtained from the a priori bit matrix.

Equation (G) can now be used to allocate the bits. It should be pointed out that in using this equation, actual energy values of the signal will be used rather than the threshold values.

Equation (G) has been simulated using the phoneme energy grouping. The bits are averaged for each energy group. The results of the simulations are shown in Ref. (10)for each of the three energy groups. Distinctly shown is a separation of the energy groups. Note that the low energy group which contains the nasalic and glide sounds is shown to separate the high energy and noise groups. This separation enhances the three-source theory of the residual signal.

It has been shown that the residual signal parallels glottal excitation (10). The use of the residual signal for encoding the speech and later exciting the speech synthesizer has several benefits. The bits are minimized in the first and second sub-bands and reduces the transmission rate for these sub-bands. It is unnecessary to transmit twice as many bits for sounds with nasalic, glide or liquid characteristics. On the other hand, the discrimination from the noise is shown to be distinct. The benefit of enhanced nasals, etc. can be seen in all sub-bands. This distinction remains clear further, because perceptual criteria will be enhanced in all sub-bands. Discrimination of sounds can be benefitted with a minimum bit allocation. The bit allocation scheme was used in the perceptual aspects of speech in sub-band coding of the prediction residual.

The cutoff frequencies for the sub-band coder are shown in Table III. The guidelines established for selection of cutoff frequencies is to represent an approximately equal contribution to the Articulation Index (2). The bands shown in Table III represent enough of the important frequencies such that intelligibility is kept.

To further explain how each band is related, the analysis of the sub-band is discussed. The sub-band coder is designed for 9600 bits/second. The transmitted coder parameters include the sub-band coded prediction residual signal, PARCOR coefficients and sync bits. Table III represents a breakdown of sub-band coder parameters for the high energy phonemes. Other parameters can be found in Reference (10).

TABLE III

SUB-BAND CODER PARAMETERS RELATIVE TO HIGH ENERGY PHONEMES

BAND	CUTOFF FREQUENCY (Hz)	SAMPLING FREQUENCY (Hz)	DECIMATION RATE	BITS ALLOCATED	TRANSMISSION RATE (b/s)
1	250-500	500	16	4.0	2000
2	500-1000	1000	8	3.0	3000
3	1142-1700	1142	7	1.5	1700
4	2000-3000	2000	4	1.0	2000
Sync and Synthesis					900
					9600 b/s

The fine tuning of quantization parameters has yet to be done. The total sub-band system requires many trade-offs in the analysis section. In the analysis section, allowance must be made for the transmission rate for each sub-band.

Signal-to-Noise Ratio Performance Measurements

In this section, performance measurements will be discussed. It has been recognized in the literature that signal-to-noise ratio (SNR) is an inadequate performance measure for speech coding (10). This inadequacy is related to the idea that additive white noise is not a good model for error waveforms in speech quantization. Generally, most authors supplement the SNR by subjective and perceptual measurements as a rule. In this work, to enhance SNR an adaptive quantizing technique is used based on the allocation of bits.

Each objective measure will be discussed next. Perhaps the most common measurement of performance is the conventional (normalized) SNR which is defined as

$$NSNR = 10 \log_{10}\left[\frac{\sum_{k=0}^{N-1} (x(k) - y(k))^2}{\sum_{k=0}^{N-1} x^2(k)}\right] \qquad (L)$$

where x(k) is the input to the coder and y(k) is the output of the decoder. It is assumed that the numerator represents the noise of the coding technique, such that as the noise decreases a smaller SNR will be the result of the summation in (L).

Another measure similar to (L) is the root-mean-square error which is defined as

$$\text{RMS SNR} = 20 \log_{10}\left[\sqrt{\frac{\sum_{n=0}^{N-1}(x(n)-y(n))^2}{N}}\right] \quad (M)$$

where x(n) and y(n) are defined as before. In (M), the error is assumed to be of random nature, and is normalized by the factor N, the number of data points.

A third measure is defined as

$$\text{MSSNR} = \frac{1}{N}\sum_{n=0}^{N-1} 10 \log\left[\frac{(x(n) - y(n))^2}{x^2(n)^2}\right] \quad (N)$$

where x(n) and y(n) are expressed as before. The representation in (N) defines an absolute error and indicates the amount of gross error only.

The results using (L), (M), and (N) are shown in Table IV. These results exemplify good coder performance. Several phonemes are used in these measurements and they give an adequate measure of the coder.

TABLE IV
SIGNAL-TO-NOISE PERFORMANCE MEASUREMENT FOR SEVERAL PHONEMES

Phoneme	RMSSNR	NSNR	MSSNR
/l/	29.2	36.7	18.2
/ɛ/	37.2	36.9	19.1
/æ/	35.1	37.4	17.5
/ʌ/	32.9	34.9	15.2
/a/	30.1	38.4	18.3
/u/	36.8	38.7	17.7
/ʃ/	29.8	38.0	18.4
/aI/	31.3	37.0	18.2
/aU/	34.3	38.4	18.7
/oU/	29.4	37.9	16.8
/eI/	33.4	39.0	17.0

CONCLUSION

The study of the energy in the prediction residual of the phonemes is shown to be a suitable excitation function rather than the conventional two-source mode. It is shown that the energy of the predication residual divides the phonemes into classes by phonemic aggregations, namely high energy, low energy, and noise groups. The high energy group includes the vowels and dipthongs. The plosive, fricative and unvoiced phonemes compose the noise group. The low energy group is represented by the glides and nasals. The bit allocation scheme discussed in this paper is based on this idea and is shown to enhance the perceptual aspects of the decoded signal. The normalization factor introduced further enhances this quality. The sub-band coder is designed to exhibit good performance in terms of signal-to-noise ratios for objective measure of quality. The signal-to-noise ratio is only an indication for quantizer performance and generally must be supplemented by subjective and perceptual measurements for speech coding and further work is necessary in this direction.

ACKNOWLEDGEMENT

I extend my sincere thanks and gratitude for invaluable advice and technical guidance to my thesis advisor, Dr. Rao Yarlagadda of Oklahoma State University.

REFERENCES

(1) Un, C.K., and D.T. Magill. "The Residual - Excited Linear Prediction Vocoder With Transmission Rate Below 9.6 Kbits/s". IEEE Transactions on Communications, Vol. COM-23, No. 12 (1975), 1466-1474.

(2) Kryter, K. "Methods for the Calculation and Use of the Articulation Index." The Journal of the Acoustical Society of America, Vol. 34, No. 11 (1962), 1689-1697.

(3) Crochiere, R.E., S.A. Webber, and J.L. Flanagan. "Digital Coding of Speech in Sub-Bands." Bell System Technical Journal, Vol. 55, No. 8 (1976) 1069-1085.

(4) Flanagan, J.L. Speech Analysis Synthesis and Perception. New York: Springer-Verlag, 1972.

(5) Rabiner, L.R., and R.W. Schafer. Digital Processing of Speech Signals. New Jersey: Prentice-Hall, 1978.

(6) Delattre, P.C., A.M. Liberman, and F.S. Cooper. "Acoustic Loci and Transitional Cues for Consonants". The Journal of the Acoustical Society of America, Vol. 27, No. 4 (1955), 769-773.

(7) Crochiere, R.E. "On the Design of Sub-Band Coders for Low-Bit-Rate Speech Communication". Bell System Technical Journal, Vol. 56, No. 5 (1977), 747-770.

(8) Kang, G.S. Application of Linear Prediction Encoding to a Narrowband Voice Digitizer. Washington, D.C.: Naval Research Laboratory, NRL Report 7774, 1974.

(9) Markel, J.D., and A.H. Gray, Jr. Linear Prediction of Speech, New York: Springer-Verlag, 1976.

(10) Burge, L.L. "Efficient Coding of the Prediction Residual". (Unpub. PhD Dissertation, Oklahoma State University, 1979).