NEUVIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 16 au 20 MAI 1983

WIDEBAND QUALITY DPCM-AQF SPEECH DIGITIZERS FOR BIT RATES OF 16-32 kb/s

C.Cengiz EVCI*,Peter J. PATRICK+ and Costas S. XYDEAS+

*T.R.T. 5, av. Réaumur-92350 LE PLESSIS-ROBINSON FRANCE
+Elec. Eng. Dept., Loughborough Uni. Leics. ANGLETERRE

RESUME

Le sujet de cet article est de présenter les résultats obtenus avec des codeurs de parole DPCM-AQF adaptatifs avec un "Voiced/Unvoiced Band Switching system" (VUBS).

Le processeur VUBS traite un signal de parole à large bande (0.3-6 kHz) et comprime ce signal de parole dans un canal téléphonique de bande 0.3-3.4 kHz. Le signal est ensuite numérisé par un codeur ADPCM utilisant un algorithme de prédiction simple mais efficace pour reproduire le signal de parole à large bande à la sortie du processeur VUBS.

Ce faisant, l'atténuation et la distorsion qui apparaît de façon systématique pendant le codage de sons non voisés du signal de parole sont réduits et donc la qualité et l'intelligibilité de ces sons non voisés sont simultanément améliorés.

Les résultats de simulation sur calculateur ainsi que les écoutes montrent que la parole reproduite avec l'association VUBS-DPCM-AQF utilisant une prédiction à corrélation est mieux appréciée que le signal de parole limité à la bande 0.3 3.4 KHz codée par les mêmes Codec pour des débits de 16 à 32 kbit/s.

SUMMARY

The aim of this paper is to present the results of the adaptive DPCM-AQF speech codecs which have been used in conjuction with a "Voiced/Unvoiced Band Switching System" (VUBS). The VUBS preprocessor operates on a wideband speech (0.3-6 kHz) and compresses the speech signal into a 0.3-3.4 kHz telephone channel. This signal is subsequently digitized by DPCM codec employing simple, but efficient prediction algorithms in order to reproduce the wideband speech from the VUBS postprocessor output.

In doing so, the attenuation and distortion that inherently exist during the encoding of unvoiced sounds of conventional telephonic speech are reduced and consequently both intelligibility and quality of unvoiced sounds can be improved.

The computer simulation results together with the informal listening experiences also indicate that the reproduced speech signal from VUBS in tandem with a DPCM-AQF codec employing a Correlation Switched Prediction, CSP, scheme is preferable to 0.3-3.4 kHz telephonic speech, digitized by the same codecs at bit rates of 16-32 kb/s.

WIDEBAND QUALITY DPCM-AQF SPEECH DIGITIZERS FOR BIT RATES OF 16-32 kb/s

I - INTRODUCTION

An efficient speech digitizer is required to possess,
a - A good speech quality at low transmission bit rate, b - A simple and therefore economical encoder and decoder design, c - Robustness to the channel impairments. However, there is at present no way of satisfying the users with all these points, and in general there must be a comprimise between the three conflicting requirements. The relative importance of these attributes depends on the application. In telephony, for example, quality and cost are the major factors in the choice of digitizers while in military applications intelligible speech quality at low bit rates is often essential.

There are two main approaches for digitization of speech signals, namely : waveform digitizers and parametric digitizers (vocoders). The concept used in vocoders and waveform coders are very different. The vocoder type of systems exploits certain properties of the speech production mechanism. Such systems extract the perceptually important features from the input speech production model is used to synthesize the speech signal. Consequently, any redundancy not effecting the perception is removed. This leads to a dramatic reduction in transmission bit rate although vocoders are highly complex and expensive systems. Also, the reproduced speech has a tendency to sound machine-like. This is mainly due to the degradation resulting from the inaccuracies of the basic model of the vocal tract and excitation processes. Only by deriving more precise models of the speech will the synthetic quality of the vocoder speech be removed (1).On the other hand, waveform techniques attempt to preserve the waveshape of the original signal. In this case, the speech signal is sampled and each sample is encoded and transmitted. As the complexity of the speech encoding systems tends to be a function of the transmitted bit rate, in contrast to vocoders, waveform coders which usually operate at higher bit rates, tend to be less complex, inexpensive and produce natural sounding speech (2).

In this paper, we concern with improving the intelligibility and quality of telephonic speech. Recalling that the unvoiced sounds such as /s/ and /f/, are usually difficult to distinguish due to speech being band limited between 0.3-3.4kHz we use the preprocessing system with speech having a bandwidth of 0.3-6.0 kHz. Our aim is to prevent the band limitation of unvoiced sounds while restricting the speech to the telephonic bandwidth. In order to achieve this a "Voiced/Unvoiced Band Switching" scheme (VUBS) is proposed (3). Such a bandwidth reduction method preprocesses the 6 kHz speech into 3.4 kHz signal that has all the necessary components for the reconstruction of wideband speech. Thus, in the content of the digital encoding we aim to maintain the bit rate while improving the speech quality. This can also be viewed as a method of bit rate reduction as the same speech quality encoding 6 kHz speech requires a larger bit rate than encoding the preprocessed 6 kHz to 3.4 kHz bandlimited speech.

In the following sections, the principle of the VUBS system is described and then, the VUBS system is connected in tandem with speech digitizers to produce a wideband (0.3-6.0 kHz) speech digitizers operating at bit rates between 16-32 kb/s. Consideration is given to digital encoding of VUBS compressed signal using DPCM speech digitizers with forward adaptive quantization (AQF) and with sequential prediction. The performance results, in terms of the segmented signal-to-noise ratio (SNRSEG) and the informal listening tests,are compared to those obtained from the digitization of 0.3-3.4 kHz band limited speech signals (BLS).

II - THE VOICED/UNVOICED BAND SWITCHING SYSTEM (VUBS)

The VUBS system offers a conceptually simple method for the transmission of relatively wideband speech,i.e. 0.3-6.0 kHz over the telephonic bandwidth, 0.3 to 3.4 kHz and appears to show an improvement in intelligibility and quality of the reconstructed speech. In telephonic bandwidth channels, certain unvoiced sounds such as /s/ or /f/ are usually perceived incorrectly because a large amount of their energy is concentrated above the upper cut-off frequency of the normal telephone channel. Therefore, by transmitting the frequency components of unvoiced speech which are perceptually most significant and still occupying a 3 kHz bandwidth, speech close in quality to the original 6 kHz speech can be perceived.

Figure 1(a) presents the block diagram of VUBS preprocessor where the 6.0 kHz speech input follows two paths ; the first path, PATH1, limits the bandwidth of speech to within the 0.3 to 3.4 kHz frequency range while in the second path, PATH2, only the 3.0 to 6.0 kHz frequency range is selected and subsequently shifted down to the 0.3 to 3.4 kHz band.The decision

concerning whether PATH1 or PATH2 is to be transmitted, is made by a voiced/unvoiced (V/UV) switch (4). This leads to the transmission of the signal in PATH1 if voiced speech is present while the signal formulated in PATH2 is transmitted when the input speech is unvoiced. The V/UV decision is also transmitted to the receiver as side information. Figure 1(b) shows the block diagram of VUBS postprocessor. When the V/UV detector indicates voiced speech to the receiver, the received signal is directed to the output via SWITCH2, if however, unvoiced speech is deemed to be present by the same detector, then the received signal is shifted up in frequency from 0.3-3.4 kHz bandwidth to 3.0-6.0 kHz range before being sent to the output. Therefore, the VUBS system recreates a signal that occupies the 6.0 kHz band, although never all of it at any instant.

III - DPCM-AQF CODERS

As consideration is given to the digital encoding of the VUBS compressed signal for bit rates between 16 and 32 kb/s, it is worthwhile to pause and recap the salient features of DPCM coders used in our simulation studies.

In DPCM-AQF systems the quantization step size Δ is adapted every W sampling instants from the input samples and is transmitted as overhead information to the receiver. Δ is defined as the quantized product of the weighted rms value of the prediction error calculated from the input samples. For this purpose, first or second-order linear predictors can be employed, i.e.,

$$\Delta = Q\left\{\alpha_1\left(\frac{1}{W}\sum_{i=2}^{W}(x_i - a_1 x_{i-1})^2\right)^{1/2}\right\} \quad (1)$$

or

$$\Delta = Q\left\{\alpha_2\left(\frac{1}{W}\sum_{i=3}^{W}(x_i - a_1 x_{i-1} - a_2 x_{i-2})^2\right)^{1/2}\right\} \quad (2)$$

where $Q\{(.)\}$ represents quantization of (.), $\{x_i\}$ is the input sequence of samples, and a_1/a_1, a_2 are the first or second-order linear prediction coefficients computed from W input samples. α_1 and α_2 are the AQF step-size constants which will be specified in the results section. The main reason of using AQF is due to its robustness to the channel errors provided that Δ is protected and correctly received.

Another important element in the coder is the predictor as it forms the error sequence to be quantized in the encoder and aids to recover the speech samples in the decoder. It usually takes the linear form,

$$y_i = \sum_{k=1}^{N} b_k \hat{x}_{i-k} \quad (3)$$

where the $\{b_k\}$, k=1,2,---,N are the prediction coefficients and $\{\hat{x}_i\}$ are the decoded speech samples.

A number of prediction schemes have been described extensively in the literature varying from the fixed to the complex sequential prediction algorithms (5,6). However, in this paper, we will confine ourselves to simple and efficient predictors having only two coefficients (N=2). An adaptive prediction algorithm which updates its $\{b_k\}$ coefficients at (i+1)th. sampling instant is obtained as :

$$b_{i+1,1} = \beta_1 b_{i,1} + \delta_1 \mathrm{sgn}(\hat{e}_i).\mathrm{sgn}(\hat{x}_{i-1}) \quad (4)$$

$$b_{i+1,2} = \beta_2 b_{i,2} + \delta_2 \mathrm{sgn}(\hat{e}_i).\mathrm{sgn}(\hat{x}_{i-2}) \quad (5)$$

where β_1, β_2 have values less than unity and δ_1, δ_2 are positive constants chosen by experiment. In Equations (4) (5), sgn(.) indicates the sign of (.) and the sequence $\{\hat{e}_i\}$ is the quantizer output as shown in Figure 2.

Furthermore, referring to the "Correlation Switched Prediction" scheme (CSP) described reference (7), in order to facilitate a faster coefficient convergent rate, we employ CSP scheme together with Equations (4)-(5). CSP scheme alters significantly the values of the prediction coefficients according to the first autocorrelation coefficient, c_1, of the input speech samples. The $\{x_i\}$ sequence is therefore divided into blocks of W samples and for each block the value of c_1 is calculated. c_1 is then compared with set of thresholds TR_j, j=1,2,3, which divide the $-1.0 < c_1 < 1.0$ range of c_1 into $(j_{max}+1) = 4$ zones. A specific zone is selected according to the value of c_1 and unique set of prediction coefficients (N=2), $[b_1', b_2']$ assigned to that zone is associated with the adaptation process of the sequential algorithm. Such scheme does not require to transmit the $[b_1', b_2']$ coefficients as these are stored in a look-uptable.

All that is required is to transmit the value of the threshold j and this can be achieved with a word having $\log_2 (j_{max} + 1) = 2$ bits. The coefficients $[b_1', b_2']$ are used as initial values of the sequential prediction coefficients expressed in Equations (4)-(5), for a block of W samples. However, if the value of c_1 does not change zones between the rth. and (r+1)th. blocks of input samples, the initial values of the coefficients $[b_1', b_2']$ for the $(r+1)^{th}$ block are the last values

of $[b_1, b_2]$ defined by the sequential algorithm after the processing of the rth. block of samples. Consequently, the construction of the look-up tables in an important issue as far as the system performance is concerned. The look-up tables for both Band Limited Speech Signal (BLS) and VUBS compressed speech are formed in a similar fashion to that of reference (7) with the exception of 4 zones being selected so that the number of occurence ($N_{occ.}$) of an each zone between the desired thresholds is equal. That is, in selecting TR_j, $N_{occ.}$ for each zone is set to NB/4.0 where NB is the total number of speech blocks of W samples.

Table 1 and 2 present the thresholds and the predictor coefficients $[b_1', b_2']$ of the 4-zone secondorder CSP schemes computed for BLS and VUBS compressed signals, respectively. The value of W is 256 and both signals are sampled at 8 kHz. The block diagram of a DPCM-AQF encoder using 4-zone CSP scheme is depicted in Figure 2.

IV - EXPERIMENTAL ARRANGEMENT

The investigation of the system shown in figure 2 was made by computer simulation. The evaluation of the performance was ascertained by informal listening tests and SNRSEG (dB) values, defined as :

$$SNRSEG = \frac{1}{NB}\sum_{r=1}^{NB} 10\log_{10} \frac{\sum_{i=1}^{W} x_{h+i}^2}{\sum_{i=1}^{W}(x_{h+i} - \hat{x}_{h+i})^2} \qquad (6)$$

where

$$h = 256(r-1)$$

The following test words from a male speaker formed both BLS signal and VUBS compressed signal ; "Sister, Father, S.K. Harvey, Shift, Thick, Talk". These signals have NB = 250 blocks of speech data and were band limited within 0.3 to 3.4 kHz and sampled at 8 kHz prior to encoding process. However, in our study, SNRSEG in Equation (6) is the average over 208 blocks as 42 blocks of speech signalsconstitute the silence section of the utterance.

V - RESULTS

The error signal $\{e_i\}$ calculated as the difference between $\{x_i\}$ and $\{y_i\}$ was quantized by QE with 4,8 or 16 levels while 256 levels were used for the quantization of the step size (see quantizer QS). The step size Δ, was then determined from Equations (1)-(2). For convenience, the linear prediction coefficients $[a_1, a_2]$ to be computed per block basis (see Equation (2)) were replaced by the coefficients given in Table 1 and 2 for BLS and VUBS compressed speech, respectively. Therefore, in generating the step size, the look-up tables are utilized and an appropriate set of coefficients are selected in accordance with the first correlation coefficient of the block of 256 speech samples. The simulation results showed that the performance of AQF that adapts its step size as Equation (2) yields 1-2 dB improvement in SNRSEG of the codecs when compared to those obtained as Equation (1). Consequently, this adaptive scheme was used for the rest of the simulation. The step size optimizing parameters α_1 and α_2 are provided in Table 3 for 2,3 or 4 bits/sample quantization accuracy. As the bits per sample were varied from 2 to 4 bits, the SNRSEG (dB) values were measured for DPCM-AQF codecs, for the 3.4 kHz band limited speech and VUBS compressed 6.0 kHz to 3.4 kHz speech. The sequential prediction scheme, employed in the codecs was defined as Equations (4)-(5) and $\beta_1, \beta_2, \delta_1, \delta_2$ assumed the values of 511.0/512.0, 1023.0/1024.0, 1.0/64.0, 1.0/128.0 respectively. Selecting the SNRSEG values resulted from the digitization of BLS signals together with CSP scheme as reference values, we noted a reduction in SNRSEG for the encoding of VUBS compressed speech and entered these reductions in Table 4. From this table it is clear that for BLS signals the performance of the codecs employing the sequential prediction scheme (N=2) is inferior compared to when they employ the sequential prediction simultaneously with the CSP scheme, i.e., 1.57, 1.61 and 1.77 dB's for 2,3 and 4 bits/sample quantization. This is attributed to the speech undergoing a transition from unvoiced to voiced speech demanding that sequential prediction scheme converges rapidly to new coefficient values, convergence rate required is so rapid that it would produce unstability when encoding other segments of speech. Therefore, when the entries of Table 1 or 2 are employed as initial values of the prediction coefficients, we overcome these difficulties during the unvoiced/voiced transitions. Notice that the reduction in SNRSEG values of DPCM-AQF codecs operating on the preprocessed signal, are marginally lower than SNRSEG obtained when the input signal is speech bandlimited to telephone channel. However, informal listening experiences for 16-32 kb/s confirmed that the VUBS system followed by DPCM-AQF codecs employing CSP scheme achieves an improvement in both intelligibility and quality of unvoiced sounds when compared to BLS signal. This being the reason that the perceptually significant high frequencies

of the unvoiced sounds are missing from the recovered bandlimited speech.

The SNRSEG performance of digitizers when encoding the wideband speech (0.3-7.6 kHz) at 32kb/s is greatly inferior (by 8-9 dB) compared to when they encode VUBS compressed speech or BLS signal at the same transmission bit rate. This is because the sampling rate must be increased from 8 kHz to 16 kHz, with the result that fewer bits per sample are available and further, the higher frequency components that exist the wideband speech signal cause overload to occur more frequently in the differential encoders compared to encoding of 0.3-3.4 kHz signal.

VI - CONCLUSIONS

The VUBS system whereby voiced sounds are band limited to 0.3 to 3.4 kHz, and unvoiced sounds up to 6.0 kHz are shifted down to telephonic bandwidth, is described. An important implication of this system is that the frequency components of unvoiced sounds which are perceptually most significant, and still occupying a 3 kHz bandwidth are transmitted. Furthermore, the study of VUBS system in conjuction with DPCM-AQF codecs employing second-order sequential predictor with CSP scheme reveals that quality and intelligibility of the recovered speech from VUBS post processor are preferred over those resulted from the digitization of band limited signals by the same codecs at bit rates of 16-32 kb/s. However, both quality and intelligibility of unvoiced speech can further be improved by employing recently proposed "Adaptive Frequency Mapping system" (AFMS) (8). One suggestion for future research is to evaluate the performance results of (AFMS) pre/post processor, connected in tandem with the DPCM-AQF codecs described in this article, and is to compare with the results obtained here.

REFERENCES

1 - B.G. HASKELL and R. STEELE, "Audio and Video Bit-Rate Reduction", Proc. IEEE, Vol. 69, N°2, February 1981, pp. 252-262.

2 - J.L. FLANAGAN and others, "Speech Coding", IEEE Trans. on Comms., Vol. COM-27, N°4, April 1979, pp. 710-137.

3 - P.J. PATRICK, R. STEELE and C.S. XYDEAS, "Voiced/Unvoiced Band Switching System for Transmission of 6 kHz Speech over 3.4 kHz Telephone Channels", The Radio and Electronic Engineer, Vol. 51, N° 5, May 1981, pp. 233-235.

4 - S.G. KNORR, "Reliable Voiced/Unvoiced Decision", IEEE Trans. on Acoustic, Speech and Sign. Proc., Vol. ASSP-27, N°3, June 1979, pp. 263-267.

5 - J.D. GIBSON, "Adaptive Prediction in Speech Differential Encoding Systems", IEEE Proc., Vol. 68, N°4, April 1980, pp. 488-525.

6 - C.S. XYDEAS, C.C.EVCI and R. STEELE, "Sequential Adaptive Predictors for ADPCM Speech Encoders", IEEE Trans. on Comms., Vol. COM-30, N°8, August 1982, pp. 1942-1954.

7 - C.S. XYDEAS and C.C. EVCI, "A Comparative Study of DPCM-AQF Speech Coders for Bit Rates of 16-32 kb/s" ICASSP 1982, Vol. 3, May 3-5, 1982, PARIS-France, pp. 1680-1683.

8 - P.J. PATRICK, R.STEELE and C.S.XYDEAS, "Frequency Compression of 7.6 kHz Speech into 3.3. kHz Bandwidth" to be published in IEEE Trans. on Comms., 1983.

j	THRESHOLD TR_j	CORRELATION ZONE	COEFF. b'_1	COEFF. b'_2
1	0.8250	0.8250 -1.0	1.54	-0.71
2	0.6450	0.6450 -0.8250	0.94	-0.28
3	0.5180	0.5180 -0.6450	0.60	-0.030
		-1.0 - 0.5180	0.27	-0.14

TABLE 1

j	THRESHOLD TR_j	CORRELATION ZONE	COEFF. b'_1	COEFF. b'_2
1	0.8250	0.8250-1.0	1.56	-0.74
2	0.6200	0.6200-0.8250	0.98	-0.33
3	0.4250	0.4250-0.6200	0.54	-0.037
		-1.0 - 0.4250	0.18	-0.32

TABLE 2

STEP SIZE CONSTANTS	BITS PER SAMPLE		
	2	3	4
α_1	1.0	0.50	0.33
α_2	1.40	0.80	0.50

TABLE 3

TYPE OF PREDICTION	NUMBER OF BITS PER SAMPLE			TYPE OF INPUT
	2	3	4	
SEQUENTIAL ALGORITHM	1.57	1.61	1.77	BLS
	2.10	2.35	2.30	VUBS COMPRESSED SPEECH
SEQ. ALGO. WITH CSP	0.79	0.56	0.42	

TABLE 4

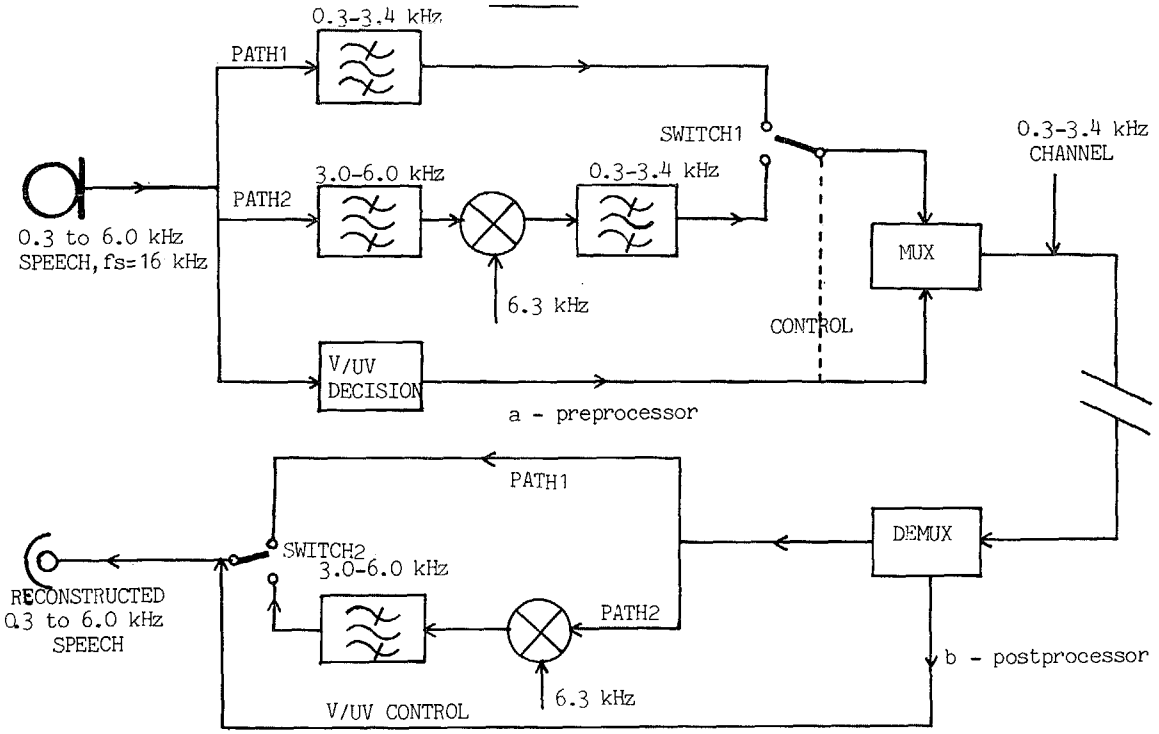

FIGURE 1 THE VUBS SYSTEM

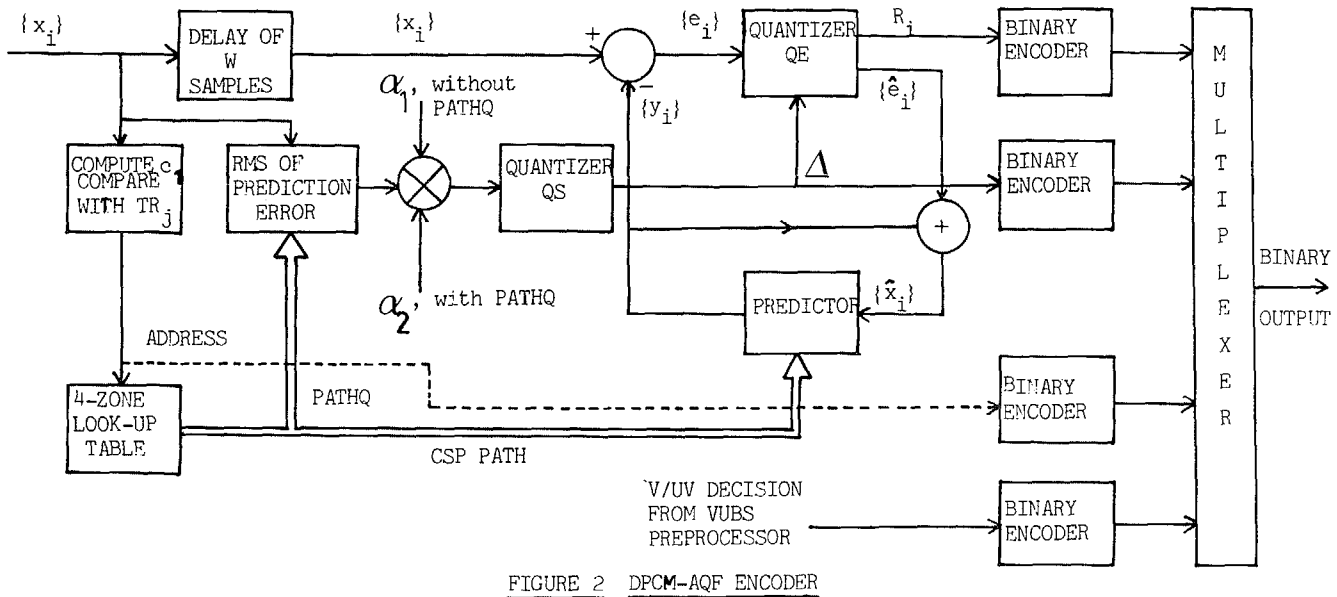

FIGURE 2 DPCM-AQF ENCODER