A LOW-RATE VOICE/VIDEO PERSONAL COMMUNICATOR SCHEME

L.Hanzo, J. Streit, R.A. Salami, *W.T. Webb †

University of Southampton, UK

Abstract

A personal communication system (PCS) transceiver is proposed and investigated. A 4.8 kbit/s transformed binary pulse excited (TBPE) linear predictive speech codec, source sensitivity matched Bose-Chaudhuri-Hocquenghem (BCH) block error correction codecs, non-coherent differentially coded non-coherent 16-level quadrature amplitude modulation (16-QAM) modem and packet reservation multiple access (PRMA) are deployed. The 2.15 kBd transceiver necessitates a signal-to-noise ratio (SNR) and signal-to-interference ratio (SIR) in excess of about 24 dB over Rayleigh-fading channels in order to support 10-11 nearly un-impaired voice conversations within a bandwith of 30 kHz. Additionally, by reserving two PRMA time slots for video telephony, an 8.52 kbps videophone user can also be supported.

1 Introduction

While the second generation digital mobile radio systems are being deployed in Europe, throughout the Pacific Rim and the United States, researchers turned their attention towards the true personal communication system (PCS) of the near future [1]. This contribution is devoted to specific algorithmic and system aspects of a complete voice/video phone transceiver suitable for the future third generation PCS. The system bandwidth was assumed to be 30 kHz, as in the American IS-54 standard system, which allowed us to assess the potential of the proposed scheme in comparison to a well known benchmarker. The system's schematic is portrayed in Figure 1, which will be described in depth throughout our further discussions.

2 The 4.8 kbit/s speech codec

In code excited linear predictive (CELP) codecs a Gaussian process with slowly varying power spectrum is used to represent the residual signal after short-term and long-term prediction, and the speech waveform is generated by filtering Gaussian excitation vectors through the time-varying linear pitch and LPC synthesis filters. The Gaussian excitation vectors of dimension N are stored in a codebook of typically 1024 entries and the optimum excitation sequence is determined by the exhaustive search of the excitation codebook. The codebook entries $c_k(n)$, $k = 1 \ldots L$, $n = 0 \ldots N-1$, after scaling by a gain factor

*Currently with Univ. of Sherbrook, Quebec, Canada
†Currently with Smith Associates, UK.

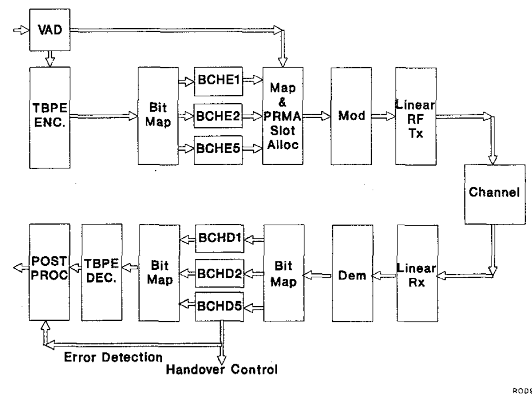

Figure 1: Transceiver schematic

G_k, are filtered through the pitch synthesis filter $1/P(z)$ and the error weighting filter $W(z) = 1/A(z/\gamma)$ to produce the weighted synthetic speech $\tilde{s}_w(n)$, which is compared to the weighted original speech $s_w(n)$.

Let $x(n)$ be the weighted original speech after removing the memory contribution of the concatenated pitch synthesis and error weighting filters $W(z) \cdot 1/P(z)$ from previous frames and $h(n)$ be the impulse response of the weighting filter $W(z)$. Then the mean squared weighted error (mswe) between the original and synthesised speech is given by:

$$E = \sum_{n=0}^{N-1} [x(n) - G_k c_k(n) * h(n)]^2. \tag{1}$$

Setting $\partial E / \partial G_k = 0$ leads to the mean square weighted error expression [2]:

$$E_{min} = \sum_{n=0}^{N-1} x^2(n) - \frac{\left[\sum_{i=0}^{N-1} \psi(i) c_k(i)\right]^2}{\sum_{i=0}^{N-1} c_k^2(i)\phi(i,i) + 2\sum_{i=1}^{N-2}\sum_{j=i+1}^{N-1} c_k(i) c_k(j) \phi(i,j)}, \tag{2}$$

where $\psi(i)$ represents the correlation between the weighting filter's impulse response $h(n)$ and the signal $x(n)$, given by $\psi(n) = x(n) * h(-n)$, while $\phi(i,j)$ represents the covariances of $h(n)$:

$$\phi(i,j) = \sum_{n=0}^{N-1} h(n-i)h(n-j). \tag{3}$$

The best innovation sequence is constituted by that codebook entry c_k with index k $(k = 1 \ldots L)$, which minimises the mean squared weighted error in *Equation 3.*

Since $\psi(n)$ and $\phi(n)$ are computed outside the error minimisation loop, the computational complexity is predetermined by the number of operations needed to evaluate the second term of *Equation 3* for all the codebook entries. For a typical excitation frame length of $N = 40$ and codebook size $L = 1024$, the CELP complexity becomes excessively high for real-time implementation. In recent years a plethora of efficient solutions have been suggested [4, 5, 6] in order to ease the computational load encountered, while still maintaining perceptually high speech quality.

In our PCS transceiver we favour a TBPE speech codec [4]. The great attraction of TBPE codecs when compared to CELP codecs accrues from the fact that the excitation optimisation can be achieved in a direct computation step [2, 4]. The sparse Gaussian excitation vector is assumed to take the form of

$$\mathbf{c} = \mathbf{A}\mathbf{b}, \tag{4}$$

where the binary vector $\mathbf{b}$ has M elements of ± 1, while the $M \cdot M$ matrix $\mathbf{A}$ represents an orthogonal transformation. Due to the orthogonality of $\mathbf{A}$ the binary excitation pulses of $\mathbf{b}$ are transformed into independent, unit variance Gaussian components of $\mathbf{c}$. The set of 2^M binary excitation vectors gives rise to 2^M Gaussian vectors of the original CELP codec.

The direct excitation computation of the TBPE codec accrues from the matrix representation of *Equation 3* using *Equation 4*, viz.:

$$E = \mathbf{x}^T\mathbf{x} - \frac{(\mathbf{\Psi}^T\mathbf{A}\mathbf{b})^2}{\mathbf{b}^T\mathbf{A}^T\mathbf{\Phi}\mathbf{A}\mathbf{b}}. \tag{5}$$

The denominator in *Equation 5* is nearly constant over the entire codebook and hence plays practically no role in the excitation optimisation. This is due to the fact that the autocorrelation matrix $\mathbf{\Phi}$ is strongly diagonal, since the impulse response $h(n)$ decays sharply. Due to the orthogonality of $\mathbf{A}$ we have $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, causing the denominator to become constant.

Closer scrutiny of *Equation 5* reveals that its second term reaches its maximum if the binary vector element is given by $b(i) = -1$, whenever the vector element $\mathbf{\Psi}^T\mathbf{A}$ is negative, and vica-versa, i.e., $b(i) = +1$ if $\mathbf{\Psi}^T\mathbf{A}$ is positive. The numerator of *Equation 5* is then constituted by exclusively positive terms, i.e., it is maximum, and the weighted mean squared error is minimum. The optimum Gaussian excitation is computed from the binary vector b using *Equation 4* in both the encoder and decoder. Only the M-bit index representing the optimum binary excitation vector $\mathbf{b}$ has to be transmitted. The evaluation of the vectors $\mathbf{\Psi}^T\mathbf{A}$ and $\mathbf{c} = \mathbf{A}\mathbf{b}$ requires $2M^2$ number of multiplications/additions, which gives typically 5 combined operations per output speech sample, a value 400 times lower than the complexity of the equivalent quality CELP codec.

The bit allocation of our TBPE codec is summarised in *Table 1*. The spectral envelope is represented by ten line spectrum frequencies (LSFs) which are scalar quantised using 36 bits. The 30 ms long speech frames having 240 samples are divided into four 7.5 ms subsegments having 60 samples. The subsegment excitation vectors $\mathbf{b}$ have 12 transformed duo-binary samples with a pulse-spacing of $D = 5$. The LTP delays (LTPD) are quantised with seven bits in odd and five bits in even indexed subsegments, while the LTP gain (LTPG) is quantised with three bits. The excitation gain (EG) factor is encoded with four bits, while the grid position (GP) of candidate excitation sequences by two bits. A total of 28 or 26 bits per subsegment is used for quantisation, which yields $36 + 2 \cdot 28 + 2 \cdot 26 = 144$ bits/30 ms, i.e., a bitrate of 4.8 kbit/s.

Parameter	Bitnumber
10 LSFs	36
LTPD	$2 \cdot 7 + 2 \cdot 5$
LTPG	$4 \cdot 3$
GP	$4 \cdot 2$
EG	$4 \cdot 4$
Excitation	$4 \cdot 12$
Total	144/20 ms

Table 1: Bit Allocation of 4.8 Kbit/s TBPE Codec

The TBPE codec was subjected to rigorous bit-sensitivity analysis [7] and the bits were assigned in three sensitivity classes for embedded source-matched forward error correction to be detailed in Section 5. Having described the proposed 4.8 kbits/s TBPE speech codec we now focus our attention on the design of the 8.52 kbps videophone codec proposed.

3 Video Codec

The video codec's outline is depicted in Figure 2. It was designed to achieve a time-invariant compression ratio associated with an encoded video rate of 8.52 kbps [1]. The codec's operation is initialised in the intra-frame mode, but once it switched to the inter-frame mode, any further mode switches are optional and only required if a drastic scene change occurs.

In the intra-frame mode the encoder transmits the coarsely quantised block averages for the current frame, which provides a low-resolution initial frame required for the operation of the inter-frame codec at both the commencement and during later stages of communications in order to prevent encoder/decoder misalignment. For 176×144 pixel CCITT standard Quarter Common Intermediate Format (QCIF) images we limited the number of video encoding bits per frame to 852. In order to transmit all block averages with a 4-bit resolution while not exceeding the 852 bits/frame rate the forced-update block size is fixed to 11 × 11 pixels.

In the motion-compensation 8×8 blocks are used. At the commencement of the encoding procedure the motion compensation (MC) scheme determines a motion vector (MV) for each of the 8×8 blocks using full-search [8]. The MC search window is fixed to 4 × 4 pels around the center of each block and hence a total of 4 bits are required for the encoding of 16 possible positions for each MV. Before the actual motion compensation takes place,

[1] The MA sequence encoded at various bitrates can be viewed under the WWW address http://rice.ecs.soton.ac.uk

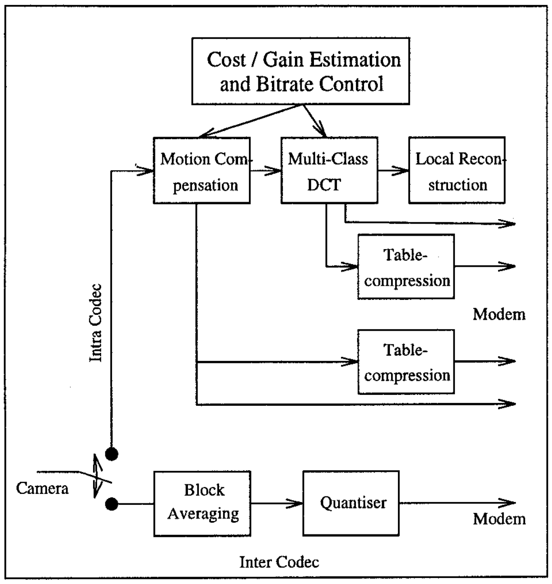

Figure 2: The intra frame codec schematic

Figure 3: PSNR versus frame index performance of the 8.52 kbps video codec for the 'Miss America' sequence

the codec tentatively determines the potential benefit of the compensation in terms of motion compensated error energy reduction. Then the codec selects those blocks as 'motion-active' whose gain exceeds a certain threshold. This method of classifying the blocks as motion-active and motion-passive results in an active/passive table, which consists of a one bit flag for each block, marking it as passive or active. In case of 8×8 blocks and 176×144 pel images this table consists of 396 entries which is compressed using elements of a two stage quad tree as follows.

First the whole table is grouped in 2×2 blocks and a four bit symbol is allocated to those blocks which contain at least one active flag. These symbols are then run length encoded and transmitted to the decoder. This concept requires a second active table containing $396 / 4 = 99$ flags in order to determine which of the two by two blocks contain active vectors. Three consecutive flags in this table are packetised to a symbol and then run length encoded. As a result, a typical 396-bit active/passive table containing 30 active flags can be compressed to less than 150 bits. Due to their low correlation the motion vectors themselves are not run length encoded. If at this stage of the encoding process the number of bits allocated to the compressed tables and active motion vectors exceeds half of the total number of available bits/frame, a number of blocks satisfying the motion-active criterion will be relegated to the motion-passive class. This process takes account of the subjective importance of various blocks and does not ignore motion-active blocks in the central eye and lip regions of the image, while relegating those, which are closer to the fringes of the image.

Pursuing a similar approach, gain control is also applied to the Discrete Cosine Transform (DCT) based compression [8]. Every block is DCT transformed and quantised. In order to take account of the non-stationary nature of the motion compensated error residual (MCER) and its time-variant frequency-domain distribution, four different sets of DCT quantisers were designed. The quantisation distortion associated with each quantiser is computed in order to be able to choose the best one. Ten bits are allocated for each quantiser, each of which are trained Max-Lloyd quantisers catering for a specific frequency-domain energy distribution class. All DCT blocks whose coding gain exceeds a certain threshold are marked as DCT-active resulting in a similar active/passive table as for the motion vectors. For this second table we apply the same run length compression technique, as above. Again, if the number of bits required for the encoding of the DCT-active blocks exceeds half of the maximum allowable number, blocks around the fringes of the image are considered DCT-passive, rather than those in the central eye and lip sections. If, however, the active DCT coefficient and activity-table do not fill up the fixed-length transmission burst, the thresholds for active DCT blocks is lowered and all tables are recomputed.

The encoded parameters transmitted to the decoder and also locally decoded in order to be used in future motion predictions. The video codec's PSNR versus frame index performance is shown in Figure 3, where an average PSNR of about 33.3 dB was achieved for the MA sequence.

4 Modulation Issues

In conventional mobile systems, such as the Pan-European GSM system [3] or the Digital European Cordless Telecommunications (DECT) scheme [2] constant envelope partial response Gaussian Minimum Shift Keying (GMSK) [2] is employed. Its main advantage is that it ignores any fading-induced amplitude fluctuation present in the received signal and hence facilitates the utilisation of power-efficient non-linear class-C amplification. In third generation personal communication systems however benign pico- and micro-cells will be employed, where low transmitted power and low signal dispersion are characteristic. Hence the employment of more bandwidth efficient multilevel modulation schemes becomes realistic. In fact the American and Japanese second generation digital systems have already opted for 2 bits/symbol modulation.

Multi-level modulation schemes have been considered in depth in reference [9] and in Chapters 17 and 18 we have shown that the bandwidth efficiency and minimum required signal-to-noise ratio (SNR) and Signal-to-Interference Ratio (SIR) of a modulation scheme in a given frequency re-use structure is dependent on the bit error ratio (BER) targeted. The required BER in turn is dependent on the robustness of the source codecs used. Furthermore, in indoors scenarios the partitioning walls and floors mitigate the co-channel interference and this facilitates the employment of 16-level Quadrature Amplitude Modulation (16QAM). In the proposed system we have opted for differentially detected Star 16-QAM [9].

5 Channel Coding

Both convolutional and block codes have been successfully used to combat the bursty channel errors. In the proposed system we have opted for binary BCH codes, since they exhibit reliable error detection, which is useful in controlling handovers and error concealment. A set of appropriate FEC codes is constituted by the BCH5=BCH(63,36,5), BCH2=BCH(63,51,2) and the BCH1=BCH(63,57,1) codes, correcting 5, 2 and 1 bits per 63-bit frame, respectively. Accordingly, the most sensitive class 1 (C1) 36 speech bits are protected by the powerful BCH(63,36,5) code, while the less vulnerable 51 class 2 (C2) and 57 class 3 (C3) bits are encoded by the BCH(63,51,2) and BCH(63,57,1) codes, respectively. The total number of protected bits is 144. The packet header conveying control information is also BCH(63,36,5) coded, hence $4 \cdot 63 = 252$ bits per 30 ms are transmitted. After adding six ramp-symbols in order to assist the transceiver in its attempt to mitigate spurious adjacent channel emissions the total bit rate becomes 8.6 kbit/s, yielding a signalling rate of 2.15 kBd.

The above FEC scheme has the advantage of curtailing BCD decoding error propagation across speech frame boundaries, although the speech codec's memory will inherently propagate errors. For a propagation frequency of 1.9 GHz, as in the future PCS the wavelength is about 15 cm, and therefore interleaving over an interval of about 40 cm ensures adequate error randomisation for the FEC scheme to work efficiently. However, for pedestrians PSs there is a danger of idling in deep fades, in which case a switch-diversity scheme is essential.

The 852 bit frame is encoded using 12 BCH(127,71,9) code words, yielding a total of 1524 bits. A pair of such codewords form a video packet of 254 bits, which is expanded by four ramp symbols to deliver a 258-bit/30 ms video packet at a corresponding rate of 2.15 kBd. Six such video packets are needed to deliver the 1524-bit BCH-coded video frame, but during the 90 ms video frame repetion time there are only three 30 ms PRMA frames. This implies that two reserved time-slots per PRMA frame are required for video users. This is equivalent to a video signalling rate of $2 \times 2.15 = 4.3$ kBd. Having resolved the choice of FEC codecs let us now consider how PRMA can be used to maximise the number of users supported.

PRMA parameter	
Channel rate	20 kBd
Speech rate	2.15 kBd
Video rate	4.3 kBd
Frame duration	30 ms
Total no. of slots	9
No. of PRMA slots	7
No. of TDMA slots	2
Slot duration	3.33 ms
Header length	63 bits
Maximum speech delay	32 ms
Speech perm. prob.	0.6

Table 2: Summary of PRMA/TDMA parameters

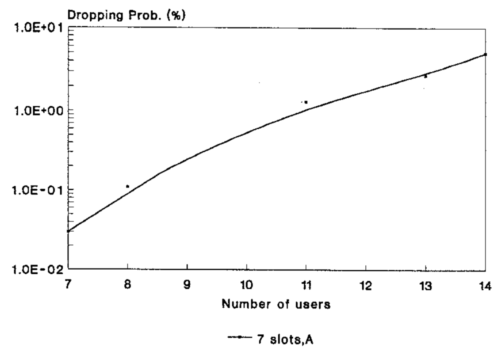

Figure 4: Packet dropping versus number of speech users performance

6 Packet Reservation Multiple Access

PRMA is a relative of slotted ALOHA contrived for conveying speech signals on a flexible demand basis via time division multiple access (TDMA) systems [9]. The voice activity detector (VAD) queues the active speech spurts to contend for an up-link TDMA time-slot for transmission to the BS. Inactive users' TDMA time slots are offered by the BS to other users, who become active and are allowed to contend for the un-used time slots with a less than unity permission probability. This measure prevents previously colliding users from consistently keep colliding in their further attempts to attain a time-slot reservation. If several users contend for an available slot, neither of them will be granted it, while if only one user requires the time slot, he can reserve it for its future communications. When many users are contending for a reservation, the collision probability is increased and hence a speech packet might have to contend for a number of consecutive slots, until its maximum contention delay of typically 32 ms expires. In this case the speech packet must be dropped, but the packet dropping probability must be kept below 1%, a value inflicting minimal degradation in perceivable speech quality in contemporary speech codecs.

The transmitted Baud rate of our transceiver was

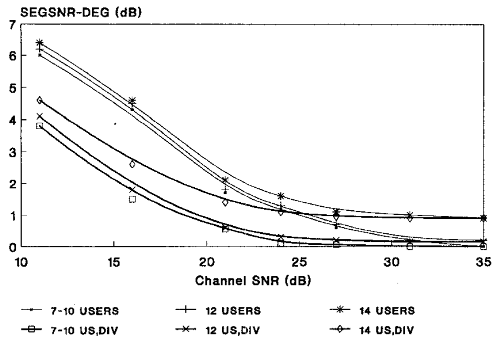

Figure 5: SEGSNR-DEG versus channel SNR performance of the proposed 16-StQAM transceiver at 30 mph with and without diversity parameterized with the number of PRMA users supported

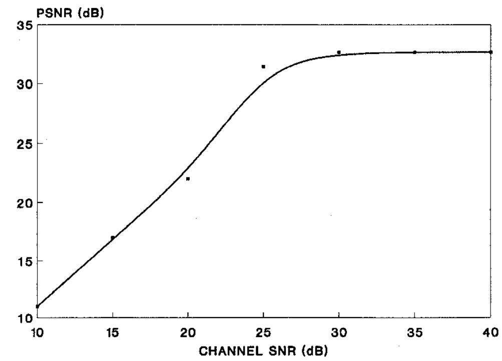

Figure 6: PSNR versus channel SNR performance of the proposed diversity-assisted videophone scheme

fixed to 20 kBd, in order for the PRMA signal to fit in a 30 kHz channel slot, as in the IS-54 system, when using a modem excess bandwidth of 50 % [9]. Hence our transceiver can accommodate TRUNC(20 kBd/2.1 kBd) = 9 time slots, where TRUNC represents truncation to the nearest integer. The slot duration was 30 ms / 9 $\approx$ 3.33 ms and one of the PRMA users was transmitting speech signals recorded during a telephone conversation, while all the other users generated negative exponentially distributed speech spurts and speech gaps with mean durations of 1 and 1.35 s. The PRMA parameters used are summarised in Table 2. Observe that two time slots are reserved for a videophone users and 7 slots are dedicated to PRMA for the speech users.

7 The Proposed PCS Transceiver

The block diagram of the proposed PCS transceiver is shown in Figure 1. The TBPE encoder outputs a 4.8 kbit/s bit stream, which is mapped in three bit sensitivity classes, C1, C2 and C3. The bits belonging to these classes are FEC encoded along with the packet header conveying control information by the BCHE1 = BCH(63,36,5), BCHE2 = BCH(63,51,2) and BCHE1 = BCH(63,57,1) encoders, respectively. The 258 bits/30ms=8.6 kbit/s FEC-coded speech packets are then transmitted at 2.15 kBd using differentially encoded 16-QAM. Only active speech spurts are queued by the VAD for transmission. If there is no other PS attempting to acquire a slot reservation, the PS is allocated this particular time slot for its future communication. In case of collision further contention is enabled with a less than unity permission probability, until either a slot is reserved or the speech packet's life-span expires. The microcellular PSs receive the slot-status near-instantaneously, implying negligible propagation delays. If this cannot be ensured, adaptive time frame alignment must be used, as proposed for the Pan-European GSM system [3].

Again, the 852 bits per frame video encoded stream is BCH(127,71,9) coded to 1524 bits/frame and transmitted at 4.3 kBd, which is equivalent to the signalling rate of two speech users. Clearly, for video telephony two time-slots are required. The video transceiver obeys the structure of Figure 1, simply the TBPE speech encoder must be replaced by the DCT video codec and no bit mapping is invoked, since a single-class BCH(127,71,9) codec is used.

The receiver seen in Figure 1 carries out the inverse functions of the transmitter. The error detection capability of the strongest BCH(63,36,5) decoder is exploited to initiate handovers and to invoke speech post-processing, if the FEC decoder happens to be overloaded due to interference or collision. The system elements of Figure 1 were simulated and the main transceiver parameters are summarised in Table 3, while the system performance will be characterised in the following Section.

8 Results and Discussion

The system performance was evaluated for the worst-case narrowband Rayleigh-fading channel characterised by a propagation frequency of 1.9 GHz, vehicular speed of 30 mph and signalling rate of 20 kBd. The packet dropping probability versus number of users curve of the proposed system is portrayed in Figure 4. Observe that about 10-11 users can be supported by our 7-slot PRMA scheme with $P_{drop} < 1\%$, a value inflicting almost negligible speech degradation.

The overall objective SEGSNR degradation (SEGSNR-DEG) versus channel SNR performance of our diversity-assisted PCS transceiver is displayed in Figure 5 parameterized with the number of PRMA users supported. While for 7-10 users no speech degradation can be observed, if the channel SNR is in excess of about 24 dB, for 12 users the SEGSNR-DEG due to PRMA packet dropping becomes noticable, although not subjectively objectionable. In case of 14 users, however, there is a consistent SEGSNR-DEG of about 1 dB due to the 4-5 % packet dropping probability seen in Figure 4. Without diversity about 5 dB higher channel SNR is necessitated in order to achieve a similar performance to that of the diversity-assisted scheme.

The PSNR versus channel SNR performance of the

1 Speech Rate (kbps)	2 PRMA Source Rate (kBd)	3 TDMA User Bandw. (kHz)	4 No of TDMA Users/Carrier	5 No of PRMA Users/Carrier	6 No of PRMA Users/slot	7 PRMA User Bandw. (kHz)	8 Min SNR and SIR (dB)
4.8	2.15	4.3	7	10	1.43	3	24

Table 3: Speech Transceiver Parameters

diversity-assisted video transceiver is portrayed in Figure 6, where in harmony with the voice transceiver a channel SNR of about 22-25 dB is required for near-unimpaired video quality. Without diversity the video scheme lacks robustness, since the corrupted run-length coded activity tables affect the whole of each video frame.

9 Summary and Conclusions

The potential of a bandwidth-efficient 2.15 kBd PRMA-assisted TBPE/BCH/16-QAM scheme has been investigated for deployment in the future PCS under the assumption of benign channel conditions. Within the 30 kHz IS-54 bandwith about 10 voice users plus a video telephone user can be supported, if channel SNR and SIR values in excess of about 24 dB can be maintained. The main transceiver features are summarised in Table 3. The system performance can be further improved at the cost of higher implementational complexity, when using a more sophisticated pilot symbol assisted, block-coded coherent square 16-QAM modem. Future work will be targeted at improving the speech quality, implementational complexity, bit rate, bandwidth occupancy and error resilience trade-off achieved invoking the adaptive transceiver reconfiguration algorithms we used in Chapters 13, 17 and 18 of reference [9].

10 Acknowledgement

The authors are grateful to the Education and Physical Sciences Research Council (EPSRC), Swindon, UK. for their financial support of the research project GR/J46845.

References

[1] IEEE Communication Magazine, Dec. 1992, Vol.30 No.12, Special Issue, PCS: The Second Generation, pp 30-133, Guest Editor: R. Steele

[2] R. Steele (Ed.): "Mobile Radio Communication", *IEEE Press-Pentech Publishers*, 1992

[3] L. Hanzo, J. Stefanov: The Pan-European digital cellular mobile radio system - known as GSM, Chapter 8, pp 677-768, in R. Steele (Ed.): "Mobile Radio Communication", *IEEE Press-Pentech Publishers*, 1992

[4] R.A. Salami. "Binary pulse excitation: a novel approach to low complexity CELP coding" in *Advances in speech coding,* Kluwer Academic Publishers,

[5] I.A. Gerson and M.A. Jasiuk. "Vector sum excitation linear prediction (VSELP) speech coding at 8 kbps". *Proc. ICASSP'90, Albuquerque, New Mexico, USA,* pp. 461-464, 3-6 Apr., 1990.

[6] J-P. Adoul et al. "Fast CELP coding based on algebraic codes". *Proc. ICASSP'87,* pp. 1957-1960, 1987.

[7] L. Hanzo, R. Salami, R. Steele, P.M. Fortune: "Transmission of Digitally Encoded Speech at 1.2 KBd for PCN", *IEE Proc.-I, vol. 139, no. 4,* pp. 437-447, Aug. 1992

[8] A.K. Jain: Fundamentals of Digital Image Processing, Prentice-Hall, 1989

[9] W.T. Webb, L. Hanzo: Modern Quadrature Amplitude Modulation, Principles and Applications for Fixed and Wireless Channels, Research Monograph, 1994, IEEE Press-Pentech Press