A 2.35 KBPS JOINT-DETECTION CDMA SPEECH TRANSCEIVER

F. C. A. Brooks, E. L. Kuan and L. Hanzo

Department of Electronics and Computer Science,
University of Southampton, Southampton S017 1BJ, UK.
Tel: +44-703-593 125, Fax: +44-703-594 508

Abstract - The speech encoded bits generated by the proposed 2.35kbps prototype waveform interpolated (PWI) speech coder are channel-coded using turbo codes or convolutional codes and transmitted using a Joint-detection Code-division Multiple Access (JD-CDMA) scheme. The speech codec combines Zinc basis function excitation (ZFE) and mixed-multiband excitation (MMBE), in order to model the mixed voiced / unvoiced speech segments at the highest possible quality. At the receiver, a joint detection algorithm is used to separate the information of the different up-link users. It is demonstrated that for channel Signal-to-Noise Ratios (SNR) in excess of about 9dB near-unimpaired speech quality is achieved, virtually independently of the number of users supported, as long as the number of users does not exceed half of the spreading factor. Furthermore, due to the limited acceptable turbo interleaver latency the more complex turbo coded system did not outperform the convolutionally coded scheme.

I. BACKGROUND

The standardisation of the third generation wireless systems has reached a mature state in Europe, the USA and Japan and the corresponding system developments are well under way right across the Globe. All three standard proposals are based on Wideband Code Division Multiple Access (W-CDMA), optionally supporting also joint multi-user detection in the up-link. In the field of speech and video source compression similarly impressive advances have been achieved and hence in this contribution a complete speech transceiver is proposed and its performance is quantified. Let us commence our discourse by considering the proposed 2.35 kbps speech codec.

II. THE 2.35 KBPS SPEECH CODEC

I. Codec Schematic

The 2.35kbps prototype waveform interpolated (PWI) speech coder amalgamating Zinc basis function excitation [1, 2] (ZFE) and mixed-multiband excitation (MMBE) is portrayed in Figure 1. As seen in Figure 1, linear predictive coding (LPC) analysis is employed, which is performed on the basis of 20ms frames, with the corresponding line spectrum frequencies (LSFs) quantized to 18 bits/frame using an LSF coding scheme similar to that of the G.729 ITU codec [3]. Following LPC analysis, pitch detection, voicing strength calculation and voiced-unvoiced (V/U) decisions are performed. For an unvoiced frame the Root-Mean-Square (RMS) energy of the LPC residual is determined, *allowing random Gaussian noise to be scaled appropriately and used at the decoder as the unvoiced excitation.*

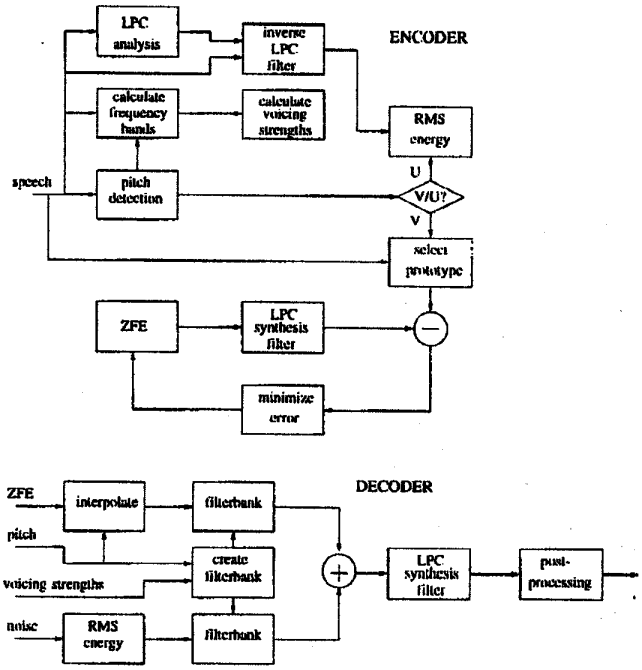

Figure 1: Schematic of the 2.35kbps codec

For a voiced speech frame initially a so-called prototype segment is selected, which is used by the codec, in order to describe a 20ms speech segment, as it will be highlighted below and also portrayed in Figure 2. This prototype segment is described by the help of a number of parameters, which will be specified during our further discourse, characterising essentially the entire voiced speech frame. Subsequently the prototype segment is passed to an analysis-by-synthesis loop, where the best voiced excitation is selected. For this voiced excitation we opted for using orthogonal zinc basis functions, with the zinc function $z(t)$ defined by Sukkar et al [1] as $z(t) = A \cdot sinc(t-\lambda) + B \cdot cosc(t-\lambda)$. As seen in Figure 3, the zinc function excitation (ZFE) exhibits a spread pulse-like shape, where the coefficients A and B describe the function's amplitude and λ defines its position.

Sukkar et al showed that this basis function is more amenable to describing the LPC residual of speech, than for example the Fast Fourier Transform (FFT). Viewing the Zinc-function's action in the time-domain, it spreads the excitation pulse's energy around the glottal closure instants and hence it provides a better replica of the LPC residual than a simple pitch-spaced pulse stream. These ZFEs are passed to the analysis-by-synthesis

VTC'99, Houston, USA

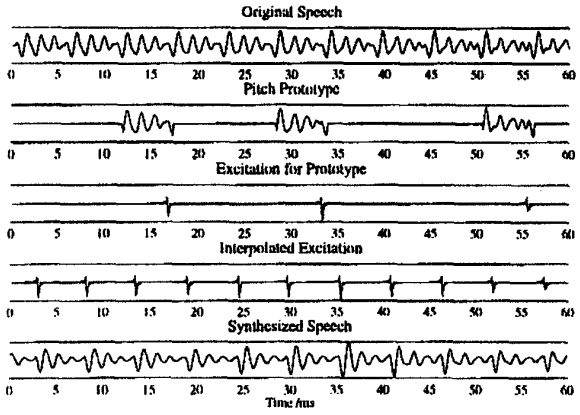

Figure 2: Example of 60 ms segments of the original speech (top to bottom), the pitch prototype and its Zinc-model as well as the interpolated excitation and the synthesized speech for a voiced utterance by a female speaker uttering /ɔ/ in 'dog'.

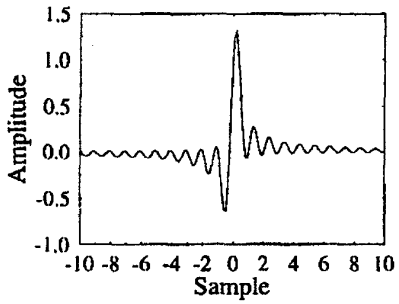

Figure 3: Typical shape of a Zinc basis function, using the expression $z(n) = A \cdot sinc(n-\lambda) + B \cdot cosc(n-\lambda)$.

loop, in order to determine the best ZFE for each prototype segment of voiced speech, a technique proposed by Hiotakakos and Xydeas [2]. The A and B ZFE parameters are then quantized and they are passed to the decoder.

Separating the speech waveform into distinct frames of either purely voiced or purely unvoiced speech usually results in synthetic speech quality due to a distortion generally termed as 'buzziness'. This distortion is particularly apparent in portions of speech that have dominant voicing in some frequency regions, but dominant noise in other frequency bands of the spectrum. Mixed-multiband excitation [5] schemes address the problem of 'buzziness' by splitting the speech into several frequency bands. These frequency bands have their voicing strength assessed individually and the corresponding excitation pattern filtered through the LPC filter is generated by superimposing the appropriately weighted Zinc-pulse and noise-based excitation sources on a subband-by-subband basis.

At the decoder seen in Figure 1, for voiced frames the ZFEs of the prototype segments are reconstructed from their A and B parameters and the excitation pulses of the 20ms speech segment, which were not transmitted are re-inserted at the positions corresponding to the pitch, while their amplitude is determined by a simple linear ramping function smoothly interpolating between the consecutive pitch prototype segments. The

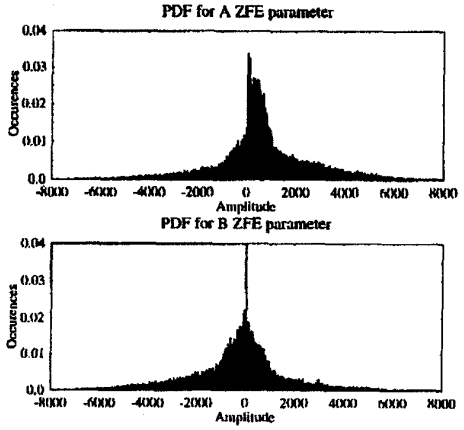

Figure 4: Typical PDF of the A and B ZFE Parameters, created from from 8 mins of BBC Radio 4's Book at Bedtime.

Quant. Scheme	SNR /dB for A	SNR /dB for B
4-bit	10.45	10.67
5-bit	18.02	19.77
6-bit	26.47	27.07

Table 1: SNR Values for Scalar Quantization of the A and B ZFE Parameters.

filter-bank stages used by the multiband excitation process have to have a bandwidth, which is an integer multiple of the pitch frequency, in order to host an integer number of pitch harmonics. Hence the transmitted pitch frequency is used to create the synthesis filterbank, in order to produce the required mixture of voiced and unvoiced excitation in all frequency bands. The resultant excitation is passed through the LPC synthesis filter, in order to produce the synthesized speech waveform, which is also passed through the adaptive postfilter [4]. Finally the waveform is passed through the pulse dispersion filter [5] in the post-processing stage of Figure 1. For further details considering the proposed speech codec the interested reader is referred to [6, 6].

II. The Speech Codec's Bit Allocation

The codec's bit allocation is summarized in Table 2, where 18 bits were reserved for LSF vector-quantization covering the groups of LSF parameters L0, L1, L2 and L3, where we used the nomenclature of the G.729 codec [3]. A one-bit flag was used for the V/U classifier, while for unvoiced speech the RMS parameter was scalar quantized with 5-bits. For voiced speech the pitch-delay was restricted to $20 \rightarrow 147$ samples, thus requiring 7-bits for transmission. The ZFE amplitude parameters A and B were scalar quantized using 6-bits. This requires the knowledge of their Probability Density Function (PDF), which is portrayed for a given training sequence in Figure 4. The Max-Lloyd quantizer was used to create 4,5 and 6-bit scalar quantizers for both the A and B parameters. Table 1 show the SNR values for the quantized A and B parameters for the various quantization schemes. On the basis of our subjective and objective investigations we concluded that the 6-bit quantization constitutes the best compromise in terms of bit rate and speech quality. The voicing strength for each frequency band was scalar quantized and since there were three frequency bands, a total of nine bits per 20 ms were allocated to voicing-strength quantisation. Thus the total number of bits for a 20ms

parameter	unvoiced	voiced
LSFs	18	18
v/u flag	1	1
RMS value	5	-
pitch	-	7
Zinc-function A	-	6
Zinc-function B	-	6
Voicing Strengths	-	3×3
total/20ms	26	47
bit rate	1.30kbps	2.35kbps

Table 2: Bit allocation for the speech codec.

frame became 26 or 47, yielding a transmission rate of 2.35kbps for the voice speech segments.

III. The Speech Codec's Error Sensitivity

Following the above description of the 2.35kbps speech codec we now investigate the extent of the reconstructed speech degradation inflicted by transmission errors. The error sensitivity is examined by individually corrupting each of the 47 bits detailed in Table 2 with a corruption probability of 10%. Employing a less than unity corruption probability is common practice, in order to allow the speech degradation caused by the previous corruption of a bit to decay, before the same bit is corrupted again, which emulates a practical transmission scenario realistically.

At the decoder for some of the transmitted parameters it is possible to invoke simple error checks and corrections. At the encoder isolated voiced, or unvoiced, frames are assumed to indicate a failure in the voiced-unvoiced decision and corrected, an identical process can be implemented at the decoder. For the pitch period parameter a smoothly evolving pitch track is created at the encoder by correcting any spurious pitch period values, and again, an identical process can be implemented at the decoder. Additionally, for voiced frame sequences phase continuity of the ZFE A and B amplitude parameters is maintained at the encoder, thus, if a phase change is perceived at the decoder, an error occurance is assumed and the previous frame's parameters can be repeated.

Figure 5 displays the so-called Segmental Signal-to-Noise Ratio (SEGSNR) and cepstral distance (CD) objective speech measures for a mixture of male and female speakers, having British and American accents. Observing Figure 5 it can be seen that both the SEGSNR and CD objectives measures rate the error sensitivity of the different bits similarly. The most sensitive parameter is the voiced-unvoiced flag, followed closely by the pitch bits, while the least sensitive parameters are the three voicing strengths bits of the bands $B1 - B3$, as seen in Figure 5.

III. CHANNEL CODING

In order to improve the performance of the system, channel coding was employed. Two types of error correction codes were used, namely, turbo codes and convolutional codes. Turbo coding is a powerful method of channel coding, which has been reported to produce excellent results [8]. Convolutional codes were used as the component codes for the turbo coding and the coding rate was set to $r = 1/2$. We used a 7×7 block interleaver as the turbo interleaver. The FMA1 spread speech/data burst 1 [11] was altered slightly to fit the turbo interleaver. Specifically, the two data blocks were modified to transmit 25 data symbols in the first block and 24 symbols in the second one. In order to obtain the soft-decision inputs required by the turbo decoder, the Euclidean distance between the CDMA receiver's

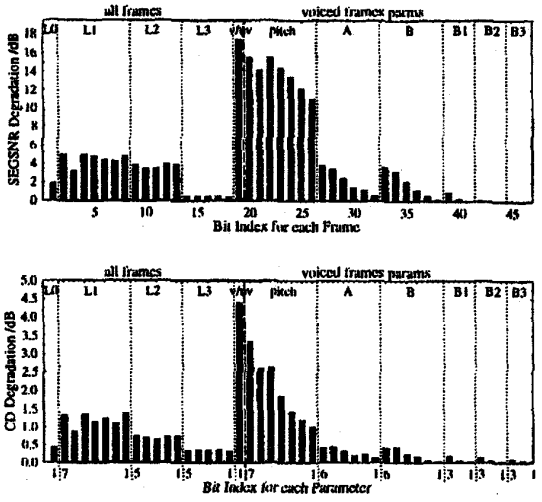

Figure 5: The error sensitivity of the different transmission bits for the 2.35kbps speech codec. For the CD degradation graph, containing the bit index for each parameter, bit 1 is the least significant bit.

data estimates and each legitimate constellation point in the data modulation scheme was calculated. The set of distance values were then fed into the turbo decoder as soft inputs. The decoding algorithm used was the Soft Output Viterbi Algorithm (SOVA) [10] with 8 iterations for turbo decoding. As a comparison, a half-rate, constraint-length three convolutional codec was used to produce a set of benchmark results. Note, however that while the turbo codec used so-called recursive systematic convolutional codecs, the convolutional codec was a non-recursive one, which has better distance properties.

IV. THE JD-CDMA SPEECH SYSTEM

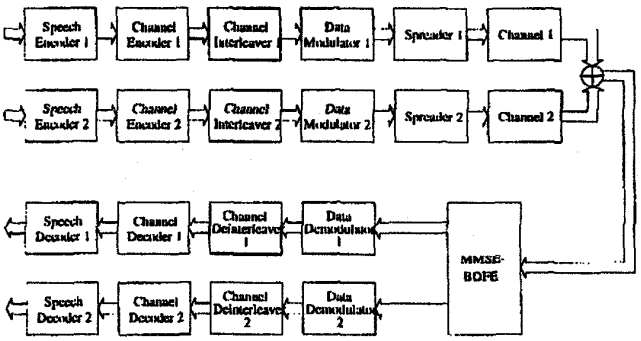

Figure 6: FRAMES-like two-user uplink CDMA system

The JD-CDMA speech system used in our investigations is illustrated in Figure 6 for a two-user scenario. The encoded speech bits generated by the 2.35kbps prototype waveform interpolated (PWI) speech codec were channel encoded using a $\frac{1}{2}$-rate turbo encoder having a frame length of 98 bits, including the convolutional codec's termination bits, where a 7×7

turbo interleaver was used. The encoded bits were then passed to a channel interleaver and modulated using 4-level Quadrature Amplitude Modulation (4-QAM). Subsequently, the modulated symbols were spread by the spreading sequence assigned to the user, where a random spreading sequence was used. The uplink conditions were investigated, where each user transmitted over a 7-path COST 207 Bad Urban channel [15], which is portrayed in Figure 7. Each path was faded independently using Rayleigh fading with a Doppler frequency of $f_D = 80$ Hz and a Baud rate of $R_b = 2.167$ MBaud. Variations due to path loss and shadowing were assumed to be eliminated by power control. The additive noise was assumed to be Gaussian with zero mean and a covariance matrix of $\sigma^2\mathbf{I}$, where σ^2 is the variance of the noise. The burst structure used in our experiments mirrored the spread/speech burst structures of the FMA1 mode of the FRAMES proposal [11]. The Minimum Mean Squared Error Block Decision Feedback Equaliser (MMSE-BDFE) was used as the multiuser receiver [13], where perfect channel estimation and perfect decision feedback were assumed. The soft outputs for each user were obtained from the MMSE-BDFE and passed to the respective channel decoders. Finally, the decoded bits were directed towards the speech decoder, where the original speech information was reconstructed.

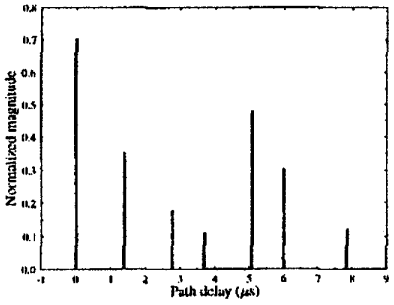

Figure 7: Normalized channel impulse response for a seven path Bad Urban channel [15].

V. SYSTEM PERFORMANCE

The BER performance of the proposed system is presented in Figures 8 and 9. Specifically, Figure 8 portrays the BER performance of a two-user JD-CDMA speech transceiver. Three different sets of results were obtained for the uncoded, turbo-coded and non-systematic convolutional-coded systems, respectively. As it can be seen from the Figure, channel coding substantially improved the BER performance of the system. However, in comparing the BER performances of the turbo-coded system and the convolutional-coded system, convolutional coding appears to offer a slight performance improvement over turbo coding. This can be attributed to the fact that a short turbo interleaver was used, in order to maintain a low speech delay, while the non-systematic convolutional codec exhibited better distance properties. It is well-understood that turbo codecs achieve an improved performance in conjunction with long turbo interleavers. However, due to the low bit rate of the speech codec 47 bits per 20ms were generated and hence we were constrained to using a low interleaving depth for the channel codecs, resulting in a slightly superior convolutional coding performance.

In Figure 9, the results were obtained by varying the number of users in the system between $K = 2$ and 6. The BER performance of the system degrades only slightly, when the number of users is increased. This is due to the employment of the joint detection receiver, which mitigates the effects of multiple access interference. It should also be noted that the performance of the system for $K = 1$ is also shown and the BER performances for $K = 2$ to 6 degrade only slightly from this single-user bound.

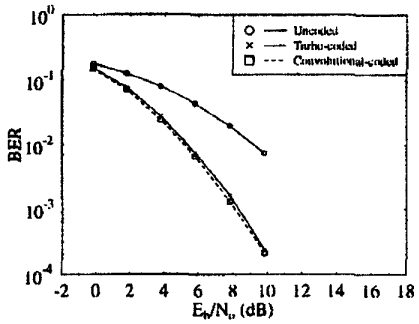

Figure 8: Comparison of the BER performance of an uncoded, convolutional-coded and turbo-coded two-user CDMA system, employing half-rate, constraint-length three constituent codes.

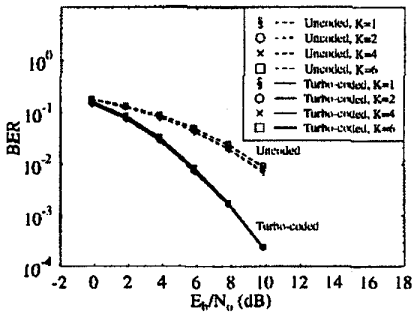

Figure 9: Comparison of the BER performance of an uncoded, convolutional-coded and turbo-coded CDMA system for $K = 2$, 4 and 6 users.

The SEGSNR and CD objective speech measures for the decoded speech bits are depicted in Figure 10, where the turbo-coded and convolutional-coded systems were compared for $K = 2$ users. As expected on the basis of our BER curves, the convolutional codecs result in a lower speech quality degradation compared to the turbo codes, which were constrained to employ a low interleaver depth. Similar findings were observed in these Figures also for $K = 4$ and 6 users. Again, the speech performance of the system for different number of users is similar, demonstrating the efficiency of the JD-CDMA receiver.

VI. CONCLUSION

The encoded speech bits generated by the 2.35kbps prototype waveform interpolated (PWI) speech codec were half-rate channel-

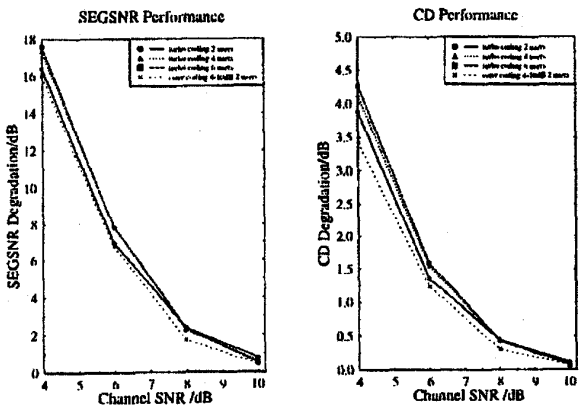

Figure 10: SEGSNR and CD objective speech measures for the decoded speech bits for $K = 2$, 4 and 6 users.

coded and transmitted using a DS-CDMA scheme. At the receiver the MMSE-BDFE multiuser joint detector was used, in order to detect the information bits, which were then channel-decoded and passed on to the speech decoder. In our work, we compared the performance of turbo codes and convolutional codes. It was shown that the convolutional codes outperformed the more complex turbo codes in terms of their BER performance and also in speech SEGSNR and CD degradation terms. This was due to the short interlever constraint imposed by the low speech delay requirement, since turbo codes require a high interlever length in order to perform effectively. It was also shown that the system performance was only slightly degraded, as the number of users was increased from $K = 2$ to 6, demonstrating the efficiency of the JD-CDMA scheme.

VII. ACKNOWLEDGMENTS

The financial support of the following organisations is gratefully acknowledged: Motorola ECID, Swindon, UK; European Community, Brussels, Belgium; Engineering and Physical Sciences Research Council, Swindon, UK; Mobile Virtual Centre of Excellence, UK.

VIII. REFERENCES

[1] R.A. Sukkar, J.L. LoCicero and J.W. Picone, "Decomposition of the LPC excitation using the zinc basis functions," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 9, pp. 1329–1341, 1989.

[2] D.J. Hiotakakos and C.S. Xydeas, "Low bit rate coding using an interpolated zinc excitation model," in *Proceedings of the ICCS 94*, pp. 865–869, 1994.

[3] CCITT, *Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic CELP*, G.729 ed., December 1995.

[4] J-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 59–70, January 1995.

[5] A.V. McCree and T.P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[6] F.C.A.Brooks, L. Hanzo: Wavelet-Based Pitch Detection for Low-rate Speech Coding, Proc. of APCC/ICCS'98, 23-27 Nov. 1998, The Westin Stamford and Westin Plaza, Singapore, pp 127-131

[7] F.C.A.Brooks, L. Hanzo: A 2.4 kbps Zinc Function Excited, Waveform Interpolated Multiband Speech Codec, Proc. of APCC/ICCS'98, 23-27 Nov. 1998, The Westin Stamford and Westin Plaza, Singapore, pp 112-117

[8] C.Berrou, A.Glavieux and P.Thitimajshima, "Near Shannon limit for error-correcting coding and decoding : turbo codes," in *Proc. of the ICC'93*, pp. 1064–1070, 1993.

[9] J.Hagenauer and P.Hoeher, "A Viterbi algorithm with soft-decision outputs and its applications," in *Proc. of IEEE GLOBECOM'89*, pp. 1680–1686, 1989.

[10] C.Berrou, P.Adde, E.Angui and S.Faudeil, "A low-complexity soft-output Viterbi decoder architecture," in *Proc. of IEEE ICC'93*, pp. 737–740, 1993.

[11] A.Klein, R.Pirhonen, J.Sköld and R.Suoranta, "FRAMES Multiple Access Mode 1 - Wideband TDMA with and without spreading," in *Proc. of the IEEE PIMRC'97*, pp. 37–41, 1997.

[12] A.Klein, G.K.Kaleh and P.W.Baier, "Zero forcing and minimum mean square error equalization for multiuser detection in code division multiple access channels," *IEEE Trans. on Vehic. Tech.*, vol. 45, pp. 276–287, May 1996.

[13] Ee-Lin Kuan, C.H. Wong, L. Hanzo: Burst-by-burst adaptive joint-detection CDMA, Proc. of IEEE VTC'99, Houston, USA

[14] W.T.Webb and L.Hanzo, *Modern Quadrature Amplitude Modulation : Principles and Applications for Fixed and Wireless Channels*. London: John Wiley and IEEE Press, 1994.

[15] Office for Official Publications of the European Communities, Luxembourg, *COST 207 : Digital land mobile radio communications, final report*, 1989.