A LOW-DELAY MULTIMODE SPEECH TERMINAL

J.P. Woodard, J.M. Torrance and L. Hanzo

Dept. of Electr. and Comp. Sc., Univ. of Southampton, SO17 1BJ, UK.
Tel: +44 17 03 59 31 25, Fax: +44 17 03 59 45 08
Email: jpw@ecs.soton.ac.uk, jmt94r@ecs.soton.ac.uk and lh@ecs.soton.ac.uk
http://rice@ecs.soton.ac.uk

ABSTRACT

The intelligent, adaptively reconfigurable wireless systems of the near future require programmable source codecs in order to optimally configure the transceiver to adapt to time-variant channel and traffic conditions. Hence we developed a programmable 8-16 kbits/s low-delay speech codec, which is compatible with the G728 16 kbits/s ITU codec [1] at its top rate and offers a graceful trade-off between speech quality and bit rate in the range 8-16 kbits/s. The issues of robustness against channel errors strongly influenced the algorithmic design of the 8-16 kbits/s speech codec, and hence special attention is devoted to these issues. Source-matched Bose-Chaudhuri-Hocquenghem (BCH) codecs combined with un-equal protection pilot-assisted 4- and 16-level quadrature amplitude modulation (4-QAM, 16-QAM) are employed in order to transmit both the 8 and the 16 kbits/s coded speech bits at a signalling rate of 10.4 kBd. In a bandwidth of 1728 kHz, which is used by the Digital European Cordless Telephone (DECT) system 55 duplex or 110 simplex time slots can be created. Good toll quality speech is delivered in an equivalent bandwidth of 15.71 kHz, if the channel signal-to-noise ratio (SNR) and signal-to-interference ratio (SIR) are in excess of about 18 and 26 dB for the lower and higher speech quality 4-QAM and 16-QAM modes, respectively.

1. INTRODUCTION

A plethora of standardised and proprietary speech codecs having fixed rates in the range of 2.4 to 64 kbits/s and using a variety of different complexity coding algorithms are available. However the intelligent adaptively reconfigurable wireless systems of the near future, studied in the framework of the European RACE and ACTS programmes, require programmable source codecs in order to optimally configure the transceiver to adapt to time-variant channel- and traffic conditions. The motivation of our work was to demonstrate the feasibility and document the performance of such a multimode voice terminal. Section 2 describes the low-delay, programmable-rate speech codec, Section 3 details the associated error sensitivity issues, while Section 4 concentrates on transmission issues. Finally, before concluding, the system performance is characterised in Section 5.

FURTHER DETAILS ON ASSOCIATED WORK CAN BE FOUND UNDER HTTP://RICE@ECS.SOTON.AC.UK.

IEEE VTC96, ATLANTA, USA

2. A LOW DELAY VARIABLE RATE SPEECH CODEC

2.1. Applications and Background

In order to assist the operation of the adaptive multimode terminal we designed a programmable 8-16 kbits/s low delay codec. There are many possible applications for such a codec. For example it could be advantageously employed in Asynchronous Transfer Mode (ATM) or Code Division Multiple Access (CDMA) networks, where the system benefits from dropping the source coding rate under heavy traffic loading or hostile channel conditions. Also it could replace the analogous but higher rate 16-40 kbits/s G726 ITU codec family and hence allow existing systems to double or triple the number of users supported.

The G728 ITU Recommendation [2], standardised in 1992, specifies a toll quality 16 kbits/s speech codec with a one way coding delay of less than 2ms. Such a low delay is achieved by using a backward adapted synthesis filter, and a frame or vector size of only 5 samples. For each 5 sample vector 10 bits are used to quantize the excitation which is fed to the 50-th order synthesis filter.

There are two alternative approaches to reducing the bit rate of the G728 codec. We can either reduce the number of bits used to encode each 5-sample original speech vector, or increase the number of speech samples per vector. Keeping the vector size fixed at five samples results in a codec with a constant delay of less than 2ms at all bit rates. However it means that at 8 kbits/s there are only 5 bits available to encode the excitation for each vector, which makes it difficult to achieve good speech quality. We will demonstrate that better reconstructed speech quality can be achieved at low rates by using 10 bits to encode the excitation for each vector, but increasing the number of speech samples per vector from 5 at 16 kbits/s to 10 for an 8 kbits/s codec. In this paper we report the results as regards to both of these approaches.

2.2. Codec Structure

Our G728-like codec is based on the Code Excited Linear Predictive (CELP) [3] speech codecs which are commonly used at low bit rates. A linear predictive synthesis filter is used to model the vocal tract of the speaker, and the excitation which drives this filter is vector quantized. For each

0-7803-3157-5/96 $5.00 © 1996 IEEE

vector the best excitation is chosen using an Analysis-by-Synthesis (AbS) approach so that the perceptually weighted error between the input speech and the reconstructed speech is minimised. The codebook indices for the best excitation are then transmitted to the decoder, where they are passed through the synthesis filter in order to produce the reconstructed speech.

In the G728 codec 10 bits are used to quantize the excitation signal, with a 3 bit scalar quantizer used to represent the excitation 'gain' and a 7 bit vector quantizer used to represent the excitation 'shape'. This split shape/gain vector quantization of the synthesis filter excitation is used to reduce to complexity of the codec, and gives very little degradation to the codec'c performance. In our variable rate codec with a constant vector size the bit rate is reduced initially by reducing the size of the gain codebook. For rates of 11.2 kbits/s and less the gain codebook is removed and the size of the shape codebook is reduced, so that at 8 kbits/s each speech vector's excitation is represented with a 5 bit index from a shape codebook. In the 8-16 kbits/s codec which increases its vector size as the bit rate is reduced we use the same 7/3 bit split shape/gain codebook structure as in G728.

An important difference between the G728 codec and most other CELP codecs is that the G728 scheme uses a backward adapted synthesis filter. This means that the filter coefficients are derived from the previously reconstructed speech, which is available at both the encoder and the decoder. Hence it is not necessary for the encoder to buffer a large (around 20ms) segment of the input speech from which to derive and transmit the predictor coefficients, and so the buffering delay of the G728 scheme is much lower than that of most CELP codecs. Furthermore, to partially compensate for the lack of Long Term Prediction (LTP) in the G728 codec a very high filter order of 50 is used, in contrast to the more conventional order of 10 used in most CELP codecs. This high order backward adapted synthesis filter is extremely important for the good performance of the G728 codec, and we found that it was almost as effective at 8 kbits/s as at 16 kbits/s. The prediction gain of the inverse synthesis filter decreased by less than 1 dB as the bit rate of our codec was reduced from 16 to 8 kbits/s.

2.3. Long Term Prediction

A long term predictor (LTP) was not used in the G728 codec because of the sensitivity of backward adapted LTP schemes to channel errors. However in applications where there is little corruption between the encoder and decoder, a LTP can give a significant improvement in the performance of the codec. Furthermore, when LTP is used, the order of the synthesis filter can be reduced to 20 with little penalty in terms of codec performance. We found that backward adaptive LTP improved the segmental SNR of the codec by around 0.5-0.75 dB. The backward adapted LTP delay was derived from the value of the pitch used in the pitch post filter, and a 3rd order predictor was used with the 3 coefficients being calculated using the autocorrelation approach on the basis of the previous excitation signal. Due to the reduction in the synthesis filter order from 50 to 20 when LTP was used, the addition of LTP had little effect on the overall complexity of the codec.

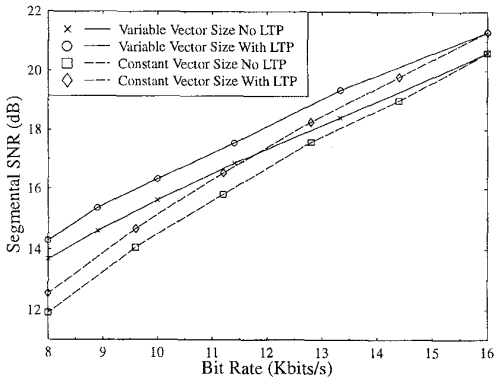

Figure 1: Segmental SNR versus Bit Rate Performance of the 8-16 kbits/s Codec With and Without LTP

2.4. Codebook Training

In the original CELP codec [3] a randomly generated Gaussian distribution was used for the excitation shape codebook, and subsequently various ternary, binary, transformed binary, algebraic and vector excited codebooks have been designed, mainly with the aim of reducing the complexity of the Analysis-by-Synthesis (AbS) procedure. However, we found that it was possible to significantly improve the performance of our variable rate codecs by carefully training gain and shape codebooks. Due to the backward adaptive nature of the codecs it was necessary to use closed loop codebook training as described in [4]. We found that trained codebooks improved the segmental SNR performance of our codecs for speakers outside the training sequence by around 1.5 dB.

2.5. Codec Performance

The segmental SNR versus bit rate performance of our 8-16 kbits/s codec is shown in Figure 1. The solid lines represent the codecs with a variable vector size, both with and without 3 tap backward adaptive LTP. The dashed lines represent the variable rate codecs with a constant vector size, again both with and without LTP. It can be seen that as expected, the variable rate codec with a constant vector size performs worse than the codec which increases its vector size. The performance gap between the two approaches increases to about 1.75 dB as the bit rate is reduced from 16 to 8 kbits/s. Furthermore, as stated above, the addition of LTP improves the performance of both codecs by between 0.5 and 0.75 dB at all bit rates. All four codecs show a graceful reduction in their segmental SNR as the bit rate is reduced, with the LTP-assisted variable vector size codec giving a segmental SNR of almost 14.5 dB at 8kbits/s.

In this section we have demonstrated the feasibility of a G728-compatible programmable 8-16 kbits/s low-delay codec. Figure 1 shows that such a codec exhibits a graceful speech quality degradation down to 8 kbits/s. In the next section we consider the robustness of the codec at 8 and 16 kbits/s to channel errors.

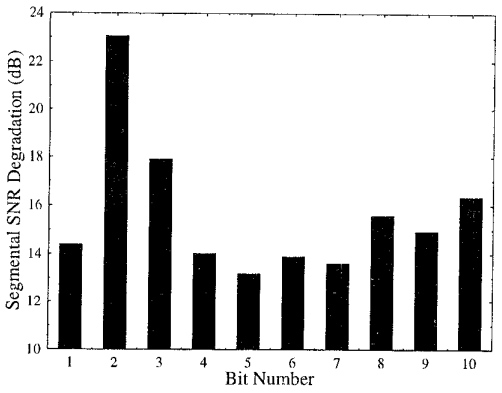

Figure 2: Degradation in G728 Segmental SNR Caused by 10 % BER in Given Bits

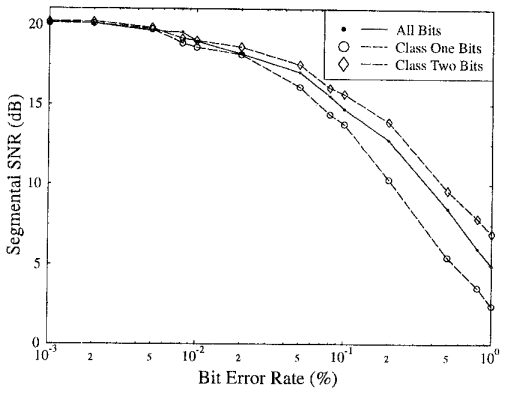

Figure 3: Segmental SNR of G728 Codec Against Channel BER

3. ERROR SENSITIVITY ISSUES

It can be seen from Figure 1 that the use of LTP in the low delay codecs increases their segmental SNR by about 0.5 to 0.75 dB. However in the presence of channel errors between the encoder and decoder the use of backward adapted LTP seriously degrades the performance of the codec. Therefore in this section we only consider the error sensitivity of the 8 and 16 kbits/s codecs without LTP. Over hostile wireless channels the system would instruct the adaptive speech codec to refrain from using LTP, whereas over benign fixed links, which are not affected by transmission errors, LTP could be used to improve the speech quality.

Figure 2 shows the sensitivity to channel errors of the ten bits used to encode each speech vector in the G728 codec. The error sensitivities were measured by, for each bit, corrupting the given bit only with a 10% Bit Error Rate (BER). This approach was taken, rather than the more usual method of corrupting the given bit in every frame, to allow account to be taken of the possible different error propagation properties of different bits [5]. Bits 1 and 2 in Figure 2 represent the magnitude of the excitation gain, bit 3 represents the sign of this gain, and the remaining bits are used to code the index of the codebook entry chosen to represent the excitation. It can be seen from this figure that not all ten bits are equally sensitive to channel errors. Notice for example that bit 2, representing the most significant bit of the excitation gain's magnitude, is particularly sensitive.

This unequal error sensitivity can also be seen from Figure 3, which shows the segmental SNR of the G728 codec for channel BERs between 0.001% and 1%. The solid line shows the performance of the codec when the errors are equally distributed amongst all ten bits, whereas the dashed lines show the performance when the errors are confined only to the 5 most sensitive bits (the so called "Class One" bits) or the 5 least sensitive bits (the "Class Two" bits). The ten bits were arranged into these two groups based on the results shown in Figure 2 – bits 2,3,8,9 and 10 formed Class One and the other five bits formed Class Two. It can be seen that the Class One bits are about two or three times more sensitive than the Class Two bits, and therefore should be more strongly protected by the error correction and modulation schemes. This aspect of the system design is considered in Section 4.

We also investigated the error sensitivity of the 8 kbits/s mode of our low delay codec. LTP was not invoked, but the codec with a vector size of ten was used because, as was seen earlier, it gave a segmental SNR almost 2 dB higher than the 8 kbits/s mode of the codec with a constant vector size of five. As discussed in Section 2.4 the vector codebook entries for our codecs were trained as described in [4]. However the 7 bit indices used to represent the 128 codebook entries are effectively randomly assigned. This assignment of indices to codebook entries does not affect the performance of the codec in error free conditions, but it is known that the robustness of vector quantizers to transmission errors can be improved by the careful allocation of indices to codebook entries [6]. This can be seen from Figure 4 which shows the segmental SNR of the 8 kbits/s codec for BERs between 0.001% and 1%. The solid line shows the performance of the codec using the codebook with the original index assignment, whereas the dashed line shows the performance of the codec when the index assignment was modified to improve the robustness of the codebook. A simple, non-optimum, algorithm was used to perform the index assignment and it is probable that the codec's robustness could be further improved by using a more effective minimisation algorithm such as simulated annealing. Also, as in the G728 codec, a natural binary code was used to represent the 8 quantized levels of the excitation gain. It is likely that the use for example of a gray code to represent the 8 gain levels could also improve the codec's robustness.

The sensitivity of the ten bits used to represent each ten speech sample vector in our 8 kbits/s codec is shown in Figure 5. Again bits 1,2 and 3 are used to represent the excitation gain, and the other 7 bits represent the index of the codebook entry chosen to code the excitation shape. As in the case of the G728 codec the unequal error resilience of different bits can be clearly seen. Note in particular how the least significant of the 3 bits representing the excitation

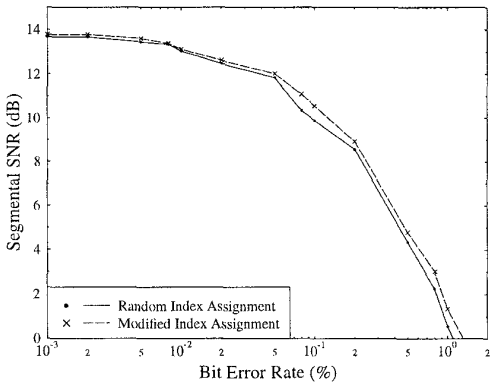

Figure 4: Segmental SNR of 8 kbits/s Codec Against Channel BER for Original and Rearranged Codebooks

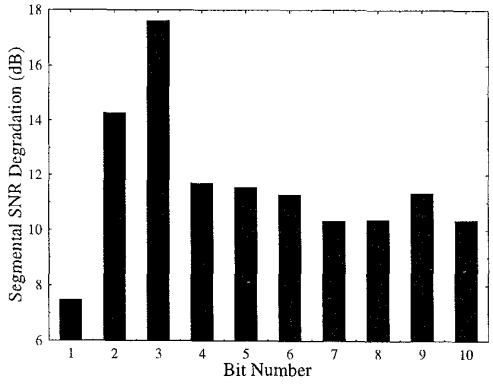

Figure 5: Degradation in 8 kbits/s Segmental SNR Caused by 10 % BER in Given Bits

gain is much less sensitive than the 7 bits representing the codebook index, but that the two most sensitive gain bits are more sensitive than the codebook index bits.

Figure 6 shows the segmental SNR of the 8 kbits/s codec for BERs between 0.001% and 1%. Again the solid line shows the performance of the codec when the errors are equally distributed amongst all ten bits, whereas the dashed lines show the performance when the errors are confined only to the 5 most sensitive Class One bits or the five least sensitive Class Two bits. The need for the more sensitive bits to be more protected by the FEC and modulation schemes is again apparent. These schemes, and how they are used to provide the required unequal error protection, is discussed in the next Section.

4. TRANSMISSION ISSUES

4.1. Higher-quality Mode

Based on the bit-sensitivity analysis presented in the previous Section we designed a sensitivity-matched transceiver scheme for both the higher and lower quality speech coding modes. Our basic design criterion was to generate an identical signalling rate in both modes in order to facilitate the transmission of speech within the same bandwidth, while providing higher robustness at a concomitant lower speech quality, if the channel conditions degrade.

Specifically, in the more vulnerable, higher-quality mode 16-level Pilot Symbol Assisted Quadrature Amplitude Modulation (16-PSAQAM) is used for the transmission of speech encoded at 16 kbps. In the more robust, lower-quality mode the 8 kbps encoded speech is transmitted using 4-PSAQAM at the same signalling rate. In our former work [5] we have found that typically it is sufficient to use a twin-class unequal protection scheme, rather than more complex multi-class arrangements. We have also shown [7] that the maximum minimum distance square 16QAM constellation exhibits two different-integrity subchannels, namely the better quality C1 and lower quality C2 subchannels, where the bit error rate (BER) difference is about a factor two in our operating Signal-to-Noise Ratio (SNR) range. This would require a forward error correction (FEC) code of twice the correction capability for achieving a similar overall performance over Gaussian channels, where the errors have a typically random, rather than bursty distribution. Over bursty Rayleigh channels an even stronger FEC code would be required in order to balance the differences between the two subchannels. After some experimentation we opted for the binary Bose-Chaudhuri-Hocquenghem BCH(127,92,5) and BCH(124,68,9) codes for the protection of the 16 kbps encoded speech bits. The weaker code was used in the lower BER C1 subchannel and the stronger in the higher BER C2 16QAM subchannel. Upon evaluating the BERs of the coded subchannels over Rayleigh channels, which are not presented here due to lack of space, we found that a ratio of two in terms of coded BER was maintained.

Since the 16 kbps speech codec generated 160 bits/10ms frame, the 92 most vulnerable speech bits were directed to the better BCH(127,92,5) C1 16QAM subchannel, while the remaining 68 bits to the other subchannel. Since the C1 and C2 sunchannels have an identical capacity, after adding some padding bits 128 bits of each subchannel were converted to 32 4-bit symbols. A control header of 30 bits was BCH(63,30,6) encoded, which was transmitted employing the more robust 4QAM mode of operation using 32 2-bit symbols. Finally, two ramp symbols were concatenated at both ends of the transmitted frame, which also incorporated four uniformly-spaced pilot symbols. A total of 104 symbols/10ms represented therefore 10 ms speech, yielding a signalling rate of 10.4 kBd. When using a bandwidth of 1728 kHz, as in the Digital European Cordless Telephone (DECT) system and an excess bandwidth of 50%, the multi-user signalling rate becomes 1152 kBd. Hence a total of INT[1152/104]=110 time-slots can be created, which allows us to support 55 duplex conversations in Time Division Duplex (TDD) mode. The timeslot duration becomes 10ms/(110 slots)$\approx$90.091μs.

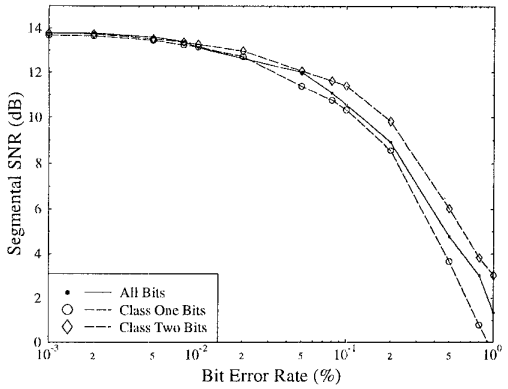

Figure 6: Segmental SNR of 8 kbits/s Codec Against Channel BER

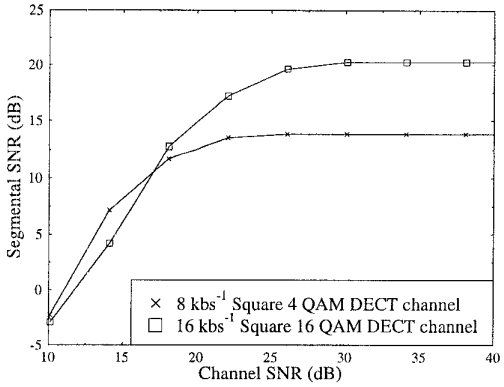

Figure 7: Segmental SNR versus Channel SNR Performance of the Proposed Multimode Transceiver

4.2. Lower-quality Mode

In the lower-quality 8 kbps mode of operation 80bits/10ms are generated by the speech codecs, but the 4QAM scheme does not have two different integrity subchannels. Here we opted for the BCH(63,36,5) and BCH(62,44,3) codes in order to provide the required integrity subchannels for the speech codec. Again, after some padding the 64-bit coded subchannels are transmitted using 2-bit/symbol 4QAM, yielding 64 symbols. After incorporating the same 32-symbol header block, 4 ramp and 4 pilot symbols, as in case of the higher-quality mode, we arrive at a transmission burst of 104 symbols/10ms, yielding an identical signalling rate of 10.4 kBd.

5. PERFORMANCE AND CONCLUSIONS

The SEGSNR versus channel SNR performance of the proposed multimode transceiver is portrayed in Figure 7 for both 10.4 kBd modes of operation. Our channel conditions were based on the DECT-like propagation frequency of 1.9 GHz, signalling rate of 1152 kBd and pedestrian speed of 1m/s=3.6 km/h, which yielded a normalised Doppler frequency of 6.3Hz/1152kBd$\approx 5.5 \cdot 10^{-3}$. Observe in the Figure that unimpaired speech quality was experienced for channel SNRs in excess of about 26 and 18 dBs in the less and more robust modes, respectively. When the channel SNR degrades substantially below 22 dB, it is more advantageous to switch to the inherently lower quality, but more robust and essentially error-free speech mode, demonstrating the advantages of the multimode concept. The effective single-user simplex bandwidth is 1728kHz/110 slots$\approx$15.71 kHz, while maintaining a total transmitter delay of 10 ms. Our current research is targeted at increasing the number of users supported using Packet Reservation Multiple Access.

6. ACKNOWLEDGEMENT

The financial support of the EPSRC, UK (GR/J46845) is gratefully acknowledged. This work also contributes towards the objectives of the Commission of European Community's (CEC) ACTS Programme under the Tasks AC410, AC411 and the award of further funding by the CEC and Motorola ECID, Swindon, UK is also thankfully acknowledged.

7. REFERENCES

[1] Juin-Hwey Chen, Richard V. Cox, Yen-Chun Lin, Nikil Jayant and Melvin J. Melchner, "A Low-Delay Coder for the CCITT 16kb/s Speech Coding Standard," *IEEE Journal on Selected Areas in Communications*, pp. 830–849, June 1992.

[2] "Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction." CCITT Recommendation G.728, 1992.

[3] Manfred R. Schroeder and Bishnu S. Atal, "Code-Excited-Linear-Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *Proceedings of the IEEE Conference on Acoustices, Speech and Signal Processing*, pp. 937–940, 1985.

[4] Juin-Hwey Chen, "High-Quality 16 Kb/s Speech Coding with a One-Way Delay Less Than 2ms," *Proceedings of the IEEE Conference on Acoustices, Speech and Signal Processing*, pp. 453–456, 1990.

[5] Lajos Hanzo and Jason P. Woodard, "An Intelligent Multimode Voice Communications System for Indoor Communications," *IEEE Transactions on Vehicular Technology*, pp. 735–748, Nov 1995.

[6] J.R.B. De Marca and N.S. Jayant, "An Algorithm for Assigning Binary Indices to the Codevectors of a Multi-Dimensional Quantizer," *Proc. of the IEEE International Conference on Communications*, pp. 1128–1132, June 1987.

[7] W. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation: Principles and Applications for Fixed and Wireless Communications.* IEEE Press-Pentech Press, 1994 ISBN 0-7803-1098-5.