A Sixth-rate, 3.8 kbps GSM-like Speech Transceiver

F.C.A Brooks, B.L Yeap, J.P Woodard and L. Hanzo

Dept. of Electronics and Computer Science,
University of Southampton, SO17 1BJ, UK.
Tel: +44-703-593 125, Fax: +44-703-594 508
Email: lh@ecs.soton.ac.uk
http://www-mobile.ecs.soton.ac.uk

Abstract

A 1.9 kbps Zinc-function excited, wave-form interpolated speech codec is proposed and its bit error sensitivity is analysed. This codec is incorporated in a standard GSM-like system, employing either convolutional or turbo coding. Ironically, the higher complexity turbo codec provides only a modest robustness gain over the standard convolutional code due to the short interleaver constraint imposed by the highly bandwidth-efficient speech codec.

1 Motivation

Although the standardisation of the third-generation European system has been completed, it is worthwhile considering potential evolutionary paths for the mature GSM system. This tendency was hallmarked by the various GSM Phase2 proposals, endeavouring to improve the services supported or by the development of the half-rate and enhanced full-rate speech codecs. In this contribution two potential improvements and their interactions in a source-sensitivity matched transceiver are considered, namely employing an approximately sixth-rate, 1.9 kbps speech codec and turbo coding.

2 The 1.9 kbps Speech Codec

For our sixth-rate GSM candidate system the 1.9kbps zinc function excited, waveform interpolated (WI) speech codec of Figure 1 is proposed, which was detailed in [2]. The sixth-rate codec operates on 20ms speech frames, where linear predictive coding (LPC) analysis is performed on each frame. The LPC coefficients are transformed to line spectrum frequencies (LSFs) and vector quantized to 18 bits/frame using an LSF coding scheme similar to that of the G.729 ITU codec[1]. Following LPC analysis, pitch detection and a voiced-unvoiced (V/U) decision are performed. For an unvoiced frame the Root-Mean-Square (RMS) energy value of the LPC residual is determined, allowing random Gaussian noise to be scaled appropriately and used at the decoder as unvoiced excitation.

The human ear has increased perceptual sensitivity to voiced speech, thus in our 1.9 kbps codec the voiced segments are more comprehensively defined than the unvoiced segments. Observing Figure 1, for a voiced speech frame initially a so-called pitch-prototype segment is selected, representing a full cycle of the pitch period. The motivation behind this is that only the pitch-prototype segment will be signalled to the decoder, where slowly evolving, seemless interpolation is used between these segments to re-instate all the non-transmitted pitch periods. The pitch-prototype segment is passed to an analysis-by-synthesis loop, where the best voiced excitation is selected under the criterion of the perceptually weighted mean-square error. For modelling this voiced excitation we opted for using the so-called orthogonal zinc basis functions, with the zinc function $z(t)$ defined by Sukkar et al [3] as $z(t) = A \cdot sinc(t-\lambda) + B \cdot cosc(t-\lambda)$. The zinc function excitation (ZFE) has a pulse-like shape, with the coefficients A and B used to describe the function's amplitude and λ defining its position. These ZFEs are passed to the analysis-by-synthesis loop to determine the best ZFE for each prototype segment of voiced speech, a technique proposed by

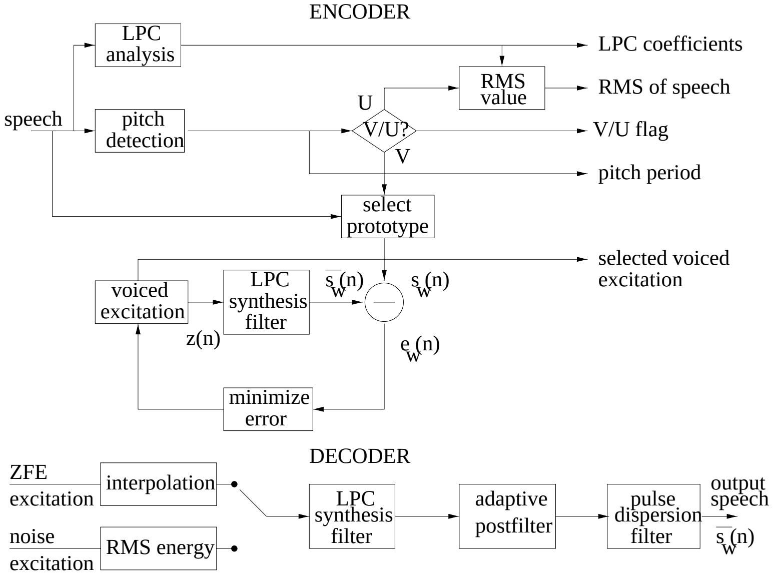

Figure 1: Schematic of the 1.9 kbps time domain prototype WI codec

parameter	unvoiced	voiced
LSFs	18	18
v/u flag	1	1
RMS value	5	-
j	3	-
pitch	-	7
A	-	6
B	-	6
total/20ms	27	38
bit rate	1.35kbps	1.90kbps

Figure 2: Bit allocation for the speech codec.

Hiotakakos and Xydeas [4]. They are then quantized and the corresponding parameters are passed to the decoder.

In low bit rate speech codecs typically the worst represented portion of speech is the rapidly evolving on-set of voiced speech. Previous speech codecs have been found to produce better quality speech by locating the emergence of voicing as precisely as possible [4, 5]. For our speech codec we identify the on-set of voicing, 'quantised or rounded' to an accuracy of the pitch duration, which is represented by the parameter j, that encodes the number of voiced speech cycles within a frame that was classified as an unvoiced one. The significance of this and the other previously introduced parameters will become more explicit in the context of the bit allocation scheme of Figure 2.

The zinc basis functions have a simple parametric representation, which at the decoder permits seamless interpolation between the prototype excitation segments, thus reinserting the excitation pulses, which were not transmitted. During the interpolation process there is no need for the location of the prototype segments to be known, since the transmitted pitch period defines the length of the interpolation process, which is constrained to be approximately 20ms. Additionally, the ZFE position parameter λ does not have to be transmitted, since the zinc pulses will be regularly spaced at pitch period intervals. The interpolated excitation is passed through the LPC synthesis filter to produce the synthesized speech waveform. Subsequently, this waveform is sent through an adaptive post-filter [6] . Finally, the waveform is passed through a pulse dispersion filter [7] to produce the output speech.

The bit allocation for the 1.9kbps speech coder is summarized in Figure 2, where 18 bits are reserved for the Line Spectral Frequency (LSF) vector-quantization covering the LSF parameter group L0, L1, L2 and L3, containing 1, 7, 5 and 5 bits, repespectively, as in G.729 [1]. A one-bit flag is used for the V/U classifier. For unvoiced speech the RMS parameter is scalar quantized with 5-bits, the j offset requires a maximum of 3-bits to encode the voiced-unvoiced transition point in terms of the number of voiced speech cycles within unvoiced frames. For voiced speech the pitch can vary from 20 → 147 samples, thus requiring 7-bits for transmission. The ZFE amplitude parameters A and B are scalar quantized with 6-bits.

3 Error Sensitivity of the Speech Codec

Following this description of the 1.9kbps speech codec we now investigate the extent of the degradation, which errors inflict upon the reproduced speech quality. The error sensitivity is examined by separately corrupting each of the 46 different bits detailed in Figure 2. Explicitly, 19 bits of these 46 different bits are sent for both V/U frames, 19 bits are sent only for V frames and 8 bits are sent only for U frames. It has been shown [8] that an error in some bits (for example the LTP bits) can produce large degradations

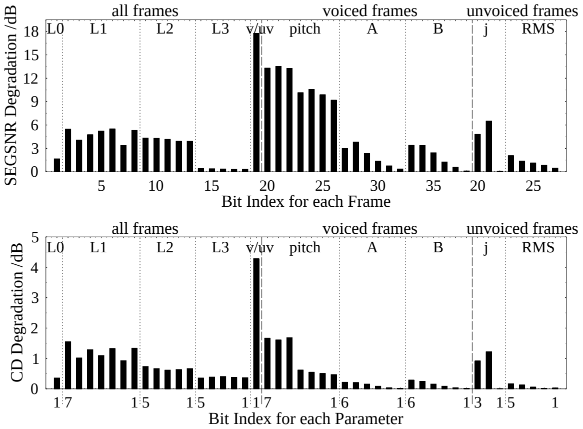

Figure 3: The error sensitivity of the different transmission bits for the 1.9kbps speech codec. The graph is divided into bits sent for all speech frames, bits sent only for voiced frames and bits sent only for unvoiced frames. For the CD degradation graph, containing the bit index for each parameter, bit 1 is the least significant bit.

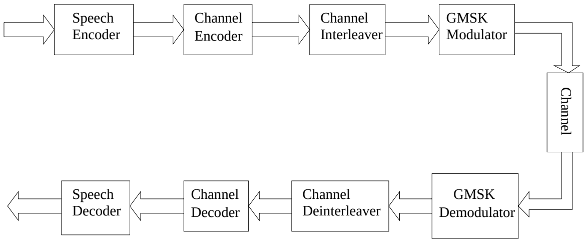

Figure 4: GSM-like system block diagram

in the speech for many subsequent frames, whereas errors in other bits (eg the LSF bits) produce little degradation in subsequent frames. To allow the different error propagation properties of different bits to affect the grade of protection different bits were given, when measuring the degradation produced by corrupting a certain bit we corrupted the given bit only in every tenth frame. This allows the affects of error propagation to die down, before the bit is corrupted again.

At the decoder for some of the transmitted parameters it is possible to make some simple error checks and corrections. At the encoder isolated voiced, or unvoiced, frames are assumed to indicate a failure in the voiced-unvoiced decision, which corrected accordingly. 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 unexpected, spurious pitch period values. Again, an identical process can be implemented at the decoder. Additionally, at the encoder for voiced frame sequences phase continuity of the ZFE A and B amplitude parameters is maintained, thus, at the decoder if a phase change occurs, an error can be assumed and the previous frame's parameter can be repeated [2].

Figure 3 displays the results for a combination of both male and female speakers, with both British and American accents. The Segmental Signal to Noise Ratio (SEGSNR) and Cepstral Distance (CD) objective speech measures were used to evaluate the speech degradations. Additionally the synthesized corrupted speech from the different bit errors were compared through informal listening tests.

Observing Figure 3 it can be seen that both the SEGSNR and CD objective measures rate the error sensitivity of the different bits similarly, both indicating that the voiced-unvoiced flag being correct is critical for successful synthesis of the output speech. This was confirmed by listening tests, which was frequently unintelligible with 10% error in the voiced-unvoiced flag bit. Additionally, from Figure 3 it can be seen that both the pitch period and the boundary shift parameter j produce a significant speech degradation due to bit errors. However, informal listening tests do not indicate such significant quality degradation, although an incorrect pitch period does produce audible distortion. It is suggested that the time misalignment introduced by the pitch period and boundary shift parameter errors is artificially increasing the SEGSNR and CD degradation values. Thus, while the SEGSNR and CD objective measures accurately show the relative sensitivities of the bits within each parameter, interpretation of the sensitivity of each parameter has to rely more on informal listening tests.

The SEGSNR and CD objective measures together with the informal listening tests allow the bits to be ordered in terms of their error sensitivities. The most sensitive bit is the voiced-unvoiced flag. For voiced frames the three most significant bits (MSB) in the LTP delay are the next most sensitive bits, followed by the four least significant LTP delay bits. For unvoiced frames the boundary parameter shift, j, is given the same protection as the most significant three pitch period bits, while the RMS value is

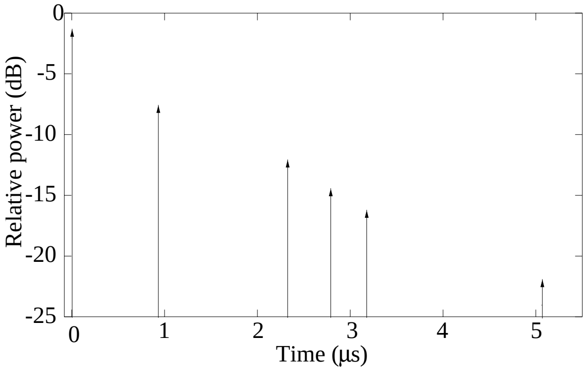

Figure 5: The impulse response of the COST207 Typical Urban channel used

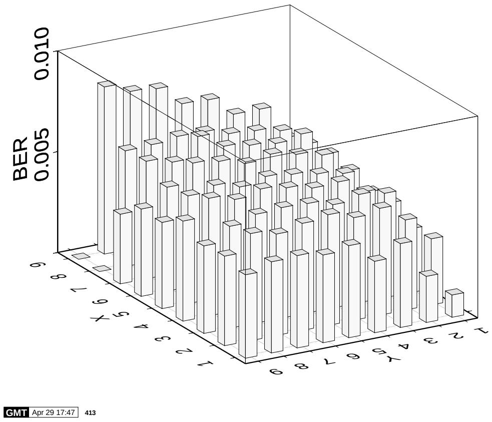

Figure 6: The error sensitivity of the different information bits within the 9x9 block interleaver used in the turbo codec

given the same protection as the group of four least significant pitch period bits and bit $A[6]$, the LSB of the ZFE amplitude A.

4 The GSM-like System

The amalgamated GSM-like system [11] is illustrated in Figure 4. In this system, the 1.9kbps speech coded bits are channel encoded with a $\frac{1}{2}$ rate convolutional or turbo encoder with an interleaving frame-length of 81 bits, including termination bits. Therefore, assuming negligible processing delay, 162 bits will be released every 40ms, or two 20ms speech frames, since the 9x9 turbo-interleaver matrix employed requires two 20ms, 38 bit, speech frames before channel encoding commences. Hence we set the data burst length to be 162 bits. The channel encoded speech bits are then passed to a channel interleaver. Subsequently, the interleaved bits are modulated using Gaussian Minimum Shift Keying (GMSK) [11] with a normalised bandwidth, $B_n = 0.3$ and transmitted at 271Kbit/s across the COST 207 [9] Typical Urban channel model. Figure 5 is the Typical Urban channel model used and each path is fading independently with Rayleigh statistics, for a vehicular speed of 50km/h or 13.89 ms^{-1} and transmission frequency of 900 MHz.

The GMSK demodulator equalises the received signal, which has been degraded by the wideband fading channel, using perfect channel estimation [11]. Subsequently, soft outputs from the demodulator are deinterleaved and passed to the channel decoder. Finally, the decoded bits are directed towards the speech decoder in order to extract the original speech information. In the following sub-sections, the channel coder and interleaver/deinterleaver, and GMSK transceiver are described.

4.1 Turbo Channel Coding

We compare two channel coding schemes, constraint-length $K = 5$ convolutional coding as used in the GSM [11] system, and a turbo channel codec. The turbo codec uses two $K = 3$ so-called Recursive Systematic Convolutionl (RSC) component codes, and 8 iterations of the Log-MAP [10] decoding algorithm. This makes it approximately 10 times more complex than the convolutional codec.

It is well known that turbo codes perform best for long interleavers. However due to the low bit rate of the speech codec we are constrained to using a low frame length in the channel codecs. A frame length of 81 bits is used, with a 9x9 block interleaver within the turbo codec. This allows two sets of 38 coded bits from the speech codec and two termination bits to be used. The BERs of the 79 transmitted bits with the 9x9 block interleaver used for the turbo codec, for a simple AWGN channel at an SNR of 2 dB, is shown in Figure 6. It can be seen that bits near the bottom right hand corner of the interleaver are

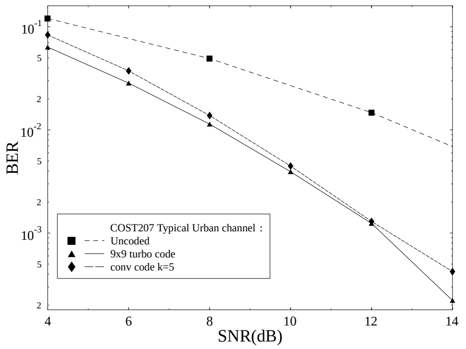

Figure 7: The BER performance for the turbo and convolutional coded systems over the COST 207 Typical Urban channel

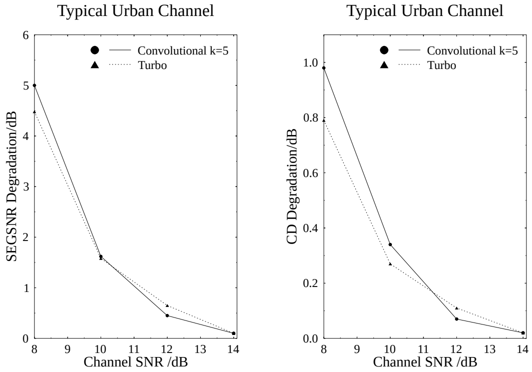

Figure 8: The speech degradation performance for the turbo and convolutional coded systems over the COST 207 Typical Urban channel

better protected than bits in other positions in the interleaver. By placing the more sensitive speech bits here we are able to give significantly more protection to the V/U flag and to some of the other sensitive speech bits, than to the low-sensitivity bits of Figure 3. Our current work investigates providing more significant un-equal error protection using turbo-codes with irregular parity bit puncturing. Lastly, an interburst channel interleaver is used, in order to disperse the bursty channel errors and to assist the channel decoders, as proposed for GSM [11].

4.2 The Turbo-coded GMSK Transceiver

As mentioned in Section 4, a GMSK modulator, with $B_n = 0.3$, which is employed in the current GSM [11] mobile radio standard, is used in our system. GMSK belongs to a class of Continuous Phase Modulation (CPM) [11], and possesses high spectral efficiency and constant signal envelope, hence allowing the use of non-linear power efficient class-C amplifiers. However, the spectral compactness is achieved at the expense of Controlled Intersymbol Interference (CISI), and therefore an equaliser, typically a Viterbi Equaliser, is needed. The conventional Viterbi Equaliser (VE) [11] performs Maximum Likelihood Sequence Estimation by observing the development of the accumulated metrics, which are evaluated recursively, over several bit intervals. The length of the observation interval depends on the complexity afforded. Hard decisions are then released at the end of the equalisation process. However, since Log Likelihood Ratios (LLRs) [12] are required by the turbo decoders, we could use a variety of soft output algorithms in place of the VE, such as the Maximum A Posteriori (MAP) [13] algorithm, the Log-MAP [10], the Max-Log-MAP [14, 15], and the Soft Output Viterbi Algorithm (SOVA) [16, 17, 18]. We chose to use the Log-MAP algorithm as it gave the optimal performance, like the MAP algorithm, but at a much lower complexity. Other schemes like the Max-Log-MAP and SOVA, are computationally less intensive, but provide sub-optimal performance. Therefore, for our work, we have opted for the Log-MAP algorithm in order to obtain the optimal performance, hence giving the upper bound performance of the system.

5 System Performance Results

The performance of our GSM-like system was compared with an equivalent conventional GSM system using convolutional codes instead of turbo codes. The $\frac{1}{2}$ rate convolutional code [11] has the same code specifications as in the standard GSM system [11]. Figure 7 illustrates the BER performance over a Rayleigh fading COST207 Typical Urban channel, and Figure 8 shows the speech degradation, in terms of both the Cepstral Distance (CD) and the Segmental SNR, for the same channel. Due to the short interleaver frame length of the turbo code the turbo- and convolutionally coded performances are fairly similar in terms of both BER and speech degradation, hence the investment of the higher complexity turbo codec is not justifiable, demonstrating an important limitation of short-latency interactive turbo-coded systems. However, we expect to see higher gains for higher bit rate speech codecs, such as for example the 260bit/20ms full-rate and the enhanced full-rate GSM speech codecs, which would allow us to use larger frame lengths for the turbo code, an issue currently investigated.

References

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

[2] F.C.A.Brooks, L. Hanzo: A 2.4 kbps Waveform Interpolation Speech Codec Incorporating Wavelet-based Techniques, submitted to IEEE Tr. on Speech and Audio Processing, 1997 [1]

[3] 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.

[4] 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.

[5] K.Yaghmaie and A.M.Kondoz, "Multiband prototype wavefrom analysis synthesis for very low bit rate speech coding," in *Proceedings of ICASSP 97*, pp. 1571–1574, 1997.

[6] 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.

[7] A.V.McCree and T.P.Barnwell III, "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.

[8] L.Hanzo and J.P. Woodard, "An Intelligent Multimode Voice Communications System for Indoor Communications," *IEEE Transactions on Vehicular Technology*, vol. 44, pp. 735–748, Nov 1995.

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

[10] Patrick Robertson, Emmanuelle Villebrun and Peter Hoeher, "A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain," *Proceedings of the International Conference on Communications*, pp. 1009–1013, June 1995.

[11] R.Steele, *Mobile radio communications.* Pentech Press, London, 1994.

[12] P.Robertson, "Illuminating the structure of code and decoder of parallel concatenated recursive systematic (turbo) codes," *IEEE Globecom*, pp. 1298–1303, 1994.

[13] L.R. Bahl, J. Cocke, F. Jelinek and J. Raviv, "Optimal Decoding of Linear Codes for Minimising Symbol Error Rate," *IEEE Transactions on Information Theory*, pp. 284–287, March 1974.

[14] W. Koch and A. Baier, "Optimum and Sub-Optimum Detection of Coded Data Disturbed by Time-Varying Inter-Symbol Interference," *IEEE Globecom*, pp. 1679–1684, Dec 1990.

[15] J.A. Erfanian, S. Pasupathy and G. Gulak, "Reduced Complexity Symbol Dectectors with Parallel Structures for ISI Channels," *IEEE Transactions on Communications*, vol. 42, pp. 1661–1671, 1994.

[16] Joachim Hagenauer, "Source-Controlled Channel Decoding," *IEEE Transactions on Communications*, vol. 43, pp. 2449–2457, Sept 1995.

[17] J. Hagenauer and P. Hoeher, "A Viterbi Algorithm with Soft-Decision Outputs and its Applications," *IEEE Globecom*, pp. 1680–1686, 1989.

[18] Claude Berrou, Patrick Adde, Ettiboua Angui and Stéphane Faudeil, "A Low Complexity Soft-Output Viterbi Decoder Architecture," *Proceedings of the International Conference on Communications*, pp. 737–740, May 1993.

[1] `http:`
`www-mobile.ecs.soton.ac.uk`